Actionable Emotion Detection in Context-aware Systems

Franci Suni Lopez

Advisor: Dr. Nelly Condori Fernandez

Committee Members:
Dr. Alex Cuadros Vargas – Universidad Católica San Pablo – Perú
Dr. Adenilso Simão – Universidade de São Paulo – Brasil
Dr. Alejandro Catalá – Universidad de Santiago de Compostela – España
Dr. José Ochoa Luna – Universidad Católica San Pablo – Perú

Thesis submitted to the
Department of Computer Science
in partial fulfillment of the requirements for the degree of
Master in Computer Science.

Universidad Católica San Pablo – UCSP
September of 2018 – Arequipa – Peru

To my parents, Margarita and Luis Enrique, who never stop giving of themselves in countless ways. To my sisters, brothers and my girldfriend, who encourage and support me.

Abbreviations

EDA Electrodermal Activity

HRV Heart Rate Variability

BVP Blood Volume Pulse

EKG Electrocardiogram

ACC Accelerometer

SAX Symbolic Aggregate approXimation

UX User Experience

SVM Support Vector Machine

TSST The Trier Social Stress Test

SC Skin Conductance

RBF Radial Basis Function

RF Random Forest

PCA Principal component Analysis

GBM Generalized Boosted Model

PPG Photoplethysmogram

Acknowledgments

First and foremost, I want to thank God for having guided me throughout these two years of study.

I would like to express my gratitude and appreciation to my advisor Nelly, for introducing me to this research area, her support throughout the whole of this work and for her advice, for the time dedicated to discussing my doubts and advances, demanding me to my utmost and patience during the preparation of this thesis.

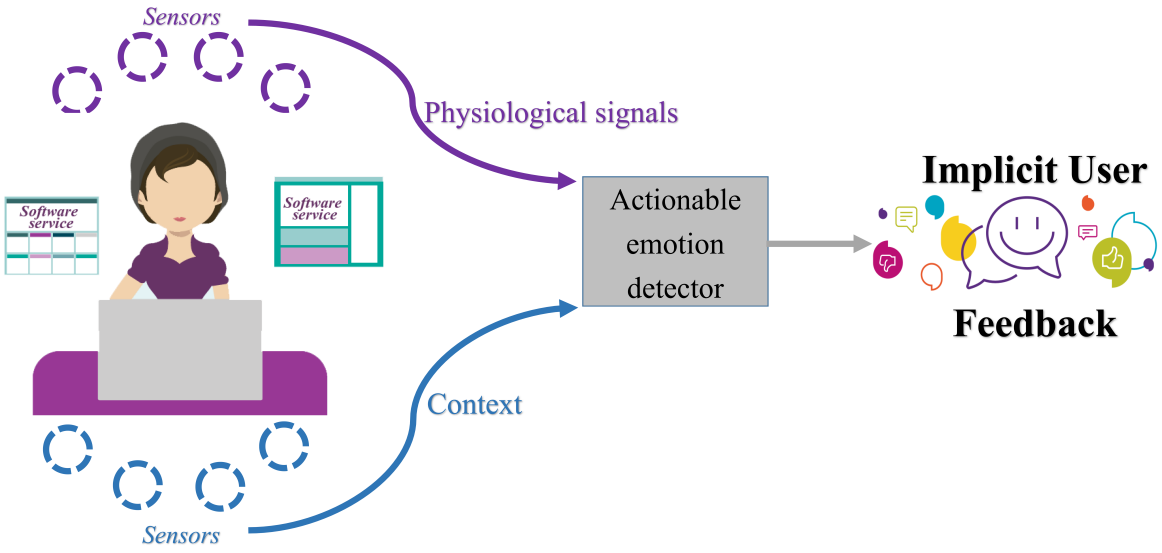
Special thanks to Patricia Lago and Nelly Condori-Fernandez for allowing me to take part in their Software and Services group (S2) and for the useful discussions we had during my stay in the Faculty of Sciences at the Vrije Universiteit in Amsterdam. To Dirk Heylen and Alejandro Catalá for facilitating me the Human Media Interaction (HMI) Lab of the University of Twente, to conduct the experiments and their early feedback.

I am going to thank all my beloved friends who directly or indirectly helped me in the preparation and presentation of this work, either by exchanging ideas, giving me advice and recommendations or by encouraging me to continue.

I want to thank my family for their love, prayers and constant support, my parents, my grandmother, sisters, bothers. I would also like to thank to my beloved girlfriend Verónica, for all her love and support. Also, and very important to me, many thanks to my best friend Luis Enrique, for supporting me and encouraging me in all the adventures undertaken, but especially for their unconditional friendship.

Finally, I would like to thank in a special way to the National Council for Science, Technology and Technological Innovation (CONCYTEC-PERU) and to the National Fund for Scientific Development, Technological and Technological Innovation (FONDECYT-CIENCIACTIVA), which through the Management Agreement 234-2015-FONDECYT have allowed the grant and financing of my studies in the Master Program in Computer Science at Universidad Católica San Pablo (UCSP).

Abstract



Ensuring the quality of user experience is very important for increasing the acceptance likelihood of software applications, which can be affected by several contextual factors that can continuously change over time (*e.g.*, emotional status of end-user). Due to these changes in the context, software continually needs to be (self-) adaptive for delivering software services that can satisfy user needs continuously. So far, online explicit user feedback has become one of the most used information sources for evaluating users' satisfaction and discovering new requirements of a given software application. However, most of these online reviews are not authenticated, and they may not always be reliable. In order to complement this explicit feedback derived from user reviews, this research proposes an approach that exploits both physiological and contextual data to be used as main inputs for detecting actionable emotions. These actionable emotions, detected during the user interaction with context-aware software applications, can be used as implicit feedback for improving the adaptability of the software and quality of the user experience. The evaluation involved in total 23 subjects in three rounds of experiments. The results of this research support the idea that emotional data expressed by users when interacting with service-based applications can be used as implicit feedback.

Keywords: implicit user feedback, actionable emotion, emotional trigger, physiological stress, user context.

Resumen

Garantizar la calidad de la experiencia del usuario es muy importante para aumentar la probabilidad de aceptación de las aplicaciones de software, las cuales pueden verse afectadas por varios factores contextuales que pueden cambiar continuamente con el tiempo (estado emocional del usuario). Debido a estos cambios en el contexto, el software continuamente debe ser auto-adaptable para entregar servicios de software que puedan satisfacer las necesidades del usuario. Hasta el momento, los comentarios explícitos en línea de los usuarios se han convertido en una de las principales fuentes de información para evaluar la satisfacción de los usuarios y descubrir nuevos requisitos de una determinada aplicación software. Sin embargo, la mayoría de estas revisiones en línea no están autenticadas y es posible que no siempre sean confiables. Con el fin de complementar esta retroalimentación explícita derivada de las reseñas de los usuarios, esta investigación propone un enfoque que utiliza los datos fisiológicos y contextuales del usuario para detectar emociones accionables. Estas emociones accionables son detectadas durante la interacción del usuario con las aplicaciones de software sensibles al contexto y pueden ser utilizadas como retroalimentación implícita para mejorar la adaptabilidad de las configuraciones del servicio software y la experiencia del usuario. La evaluación de este trabajo se basó en tres experimentos, con una población de 23 personas en total. Los resultados obtenidos respaldan la idea de que los datos emocionales expresados por los usuarios durante su interacción con aplicaciones basadas en servicios, pueden ser usado como retroalimentación implícita.

Palabras clave: retroalimentación implícita del usuario, emoción accionable, desencadenante emocional, estrés fisiológico, contexto del usuario.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and context

Nowadays, software that interacts with other software, systems, devices, sensors and with people are playing an increasingly dominant role in our lives and daily activities. Moreover, due to the continuous evolution of IT technologies such as wearable sensors, these software systems have become more complicated but at the same time adaptable. However, designing software that can detect the occurrence of changes in the context, reason about their effects, and possibly react to them in a self-adaptive manner has become a real challenge for software engineers (*e.g.,* de Lemos et al., 2017; Condori-Fernandez, 2017; Huang and Miranda, 2015; Qureshi et al., 2010). These challenges are related to the capability of changing at run-time and analyzing the user context and behavior.

To achieve the self-adaptation of a system, it is required to obtain information about any response of the interaction between the system and the user (*i.e.,* user feedback) (Ramaprasad, 1983), also the usefulness of this information depends on its consistency (it has to provide a positive or negative judgment), and its credibility and accuracy (Mezhoudi and Vanderdonckt, 2015). Due to user feedback has become a resource of valuable information for software developers, companies, service providers and users that are looking for an opinion on the quality of a service or product based on previous experiences (*i.e.,* stakeholders), it has been largely investigated by different computer science fields, especially in software engineering. According to Morales-Ramirez et al. (2015), user feedback can be classified into *implicit* when it is expressed by the behavior of the user (*e.g.,* facial expressions, time on page, click-stream, scrolling, mouse movement tracking) (*e.g.,* Peska, 2016; Jaramillo Garcia et al., 2015; Kasiran and Yahya, 2007), and *explicit* represented by filling questionnaires, sending a report about an error, suggesting new functionalities, or rating functionality or performance of software applications (*e.g.,* Kim et al., 2016; Li and Chen, 2016; Morales-Ramirez et al., 2015; Jawaheer et al., 2010).

In this context, emotions have emerged as a prominent source of information from user behavior, and it could be classified as implicit user feedback. In particular, emotion recognition from physiological data has had a significant impact on reliably, because it is based on understanding the automatic responses of the autonomic nervous system (*e.g.,* Valenza and Scilingo, 2014; Boucsein, 2012; Dawson et al., 2007; Edelberg, 1972). For instance, Condori-Fernandez and Suni Lopez (2017) introduced the idea of measuring emotions to empower the adaptability of software services at run-time. As a result of this preliminary research, authors argued that negative emotions (*i.e.,* stress) could be used as implicit user feedback for enhancing adaptability and user experience in mobile applications. However, to provide relevant user feedback about the interaction and the software service, it is needed to distinguish between emotions generated by the software service (*i.e.,* an actionable emotion) and emotions generated by others triggers (*e.g.,* environmental sounds, temperature or thoughts). An *actionable emotion* represents all those emotions that are expressed by a user within the time interval in which a software service is also delivered (Condori-Fernandez, 2017).

1.2 Problem statement

The main issue is how to collect user feedback in terms of credibility and accuracy (Mezhoudi and Vanderdonckt, 2015). In this direction, different works have been proposed, such as (Peska, 2016), (Maalej et al., 2016), (Mezhoudi and Vanderdonckt, 2015), (Morales-Ramirez et al., 2015), (Leiva, 2011), (Jawaheer et al., 2010), (Lee et al., 2008), (Claypool et al., 2001) or (Eisenstein and Puerta, 2000). Although emotions can be used to provide implicit user feedback, the recognition of human emotions still has many challenges to address. For instance, modeling the variables that intervene and influence the user context is one of these challenges. In this context, many works have been proposed (*e.g.,* Girardi et al., 2017; Mozos et al., 2017; Sriramprakash et al., 2017; Garcia-Ceja et al., 2016; Guendil et al., 2015; Sandulescu et al., 2015; Bogomolov et al., 2014; Sano and Picard, 2013; Kocielnik et al., 2013; Bauer and Lukowicz, 2012; Carneiro et al., 2012; Sun et al., 2012; Canento et al., 2011; Healey and Picard, 2005); however, most of them are not aware of possible threats from user context and they recognize emotions focused only on a predefined task. Also, none of them address emotion recognition related to the interaction with a software service (actionable emotion). In line with these notions, two challenges have been identified: changes at run-time introduce a higher degree of uncertainty (de Lemos et al., 2017); the analysis of user context (*e.g.,* thoughts, feelings, intentions) to be exploited as feedback of (self-)adaptive systems (Condori-Fernandez, 2017; Huang and Miranda, 2015) demands higher effort for gathering and aggregating data from different sources.

1.3 Goals and objectives

The main research focus lies on *detecting actionable emotions in context-aware systems, as a proposal to provide relevant and reliable implicit user feedback*. To achieve this primary goal, the following specific objectives are defined:

- Design the software architecture of an actionable emotion detector.
- Identify the most suitable emotional model for detecting emotions (stress).
- Implement a real-time stress detector that uses physiological data.
- Construct a context analyzer to interpret data from the user context.
- Define the base rules of the inference engine for determining whether detected stress is actionable.
- Evaluate the performance and accuracy of the actionable emotion detector.

1.4 Research methodology

The nature of this research lends to the use of the Design Science framework Wieringa (2014). Design science is the design and investigation of artifacts in context. In this thesis, an actionable emotion detector is designed. This methodology emphasizes the connection between knowledge and practice, demonstrating that is possible to produce scientific knowledge through the design and modeling of proposals. This framework consists of five phases:

- **Problem research.** In this phase the problem is investigated, through the solution of practical questions and knowledge questions. This interaction allows knowing in depth the problem and the state of the art on stress recognition, which are presented in Chapters 2 and 3.
- **Solution design.** After understanding the problem, this phase aims to describe and to model possible solutions. The design of the actionable detector is presented in Chapter 4.
- **Design validation.** In this phase, a prototype of the designed solution is validated concerning if the solution is closer to the goal (detection of actionable emotions). The validation of our design solution is carried out through a set of experiments reported in Chapters 5 and 6.
- **Treatment implementation.** In this phase, the actionable emotion detector is implemented according to the solution design, described in Chapter 4.

- **Implementation evaluation.** The final phase aims to evaluate the implementation of the actionable emotion detector in a real context. A case study, designed to evaluate the performance of our solution is reported in Chapter 7.

Figure 1.1 shows the design cycle; all phases are in constant interaction, which allows restructuring the solution if the results are not as expected.

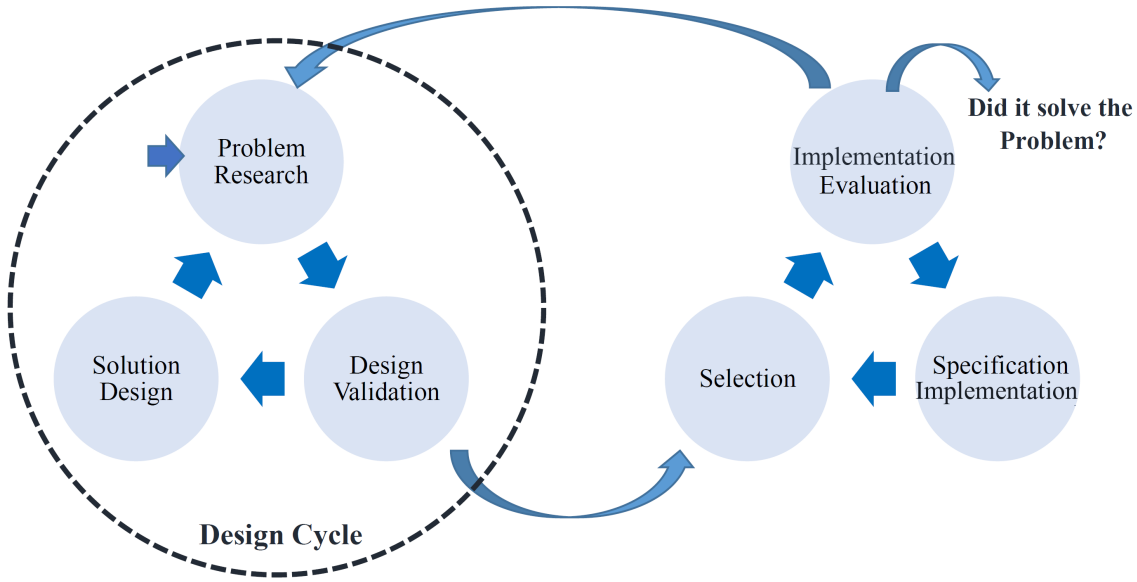


Figure 1.1: Design cycle, image extracted from (Wieringa, 2014).

1.5 Thesis outline

This thesis document is organized as follows, Chapter 2 presents the background and theory of human emotions to support the research. The literature review of user feedback and emotion recognition is shown in Chapter 3. Chapter 4 presents the software architecture and the requirements of the actionable emotion detector; also it explains the algorithms for recognizing physiological stress and for processing the user context. During this research was carried out three different experiments: Chapter 5 describes the design and results of the experiment to evaluate the stress detector; Chapter 6 presents the experiment to evaluate the performance of the actionable emotion detector, and Chapter 7 reports the design and results of a case study for evaluating the performance of the actionable emotion detector in real-life conditions. Finally, the conclusions, limitations and future works are discussed in Chapter 8.

Chapter 2

Background

This chapter presents the basic concepts and theory about emotions, which are needed to understand the proposal to detect actionable emotions. Firstly, Section 2.1 explains the theory and models of emotions. Next, theory about stress from EDA signals is detailed in Section 2.2. Finally, Section 2.3 presents concepts about emotional triggers.

2.1 Human emotions

In psychology, for decades researchers have tried to reach a common understanding of what is an emotion? or what can be considered an emotion?. Although, there is no global understanding of this term, each human has a relative knowledge of the concept, and even has the innate ability to recognize and generate these emotions physically. A general definition of "emotion" generally refers to *an affective state, often accompanied by specific physiological and mental characteristics that directly affect thoughts and behavior in humans* (Ekman, 1999). Nevertheless, this definition can be considered incomplete because it does not really cover all the scope of the emotion concept, which ranges from biological, physiological and behavioral changes to our own perception of things, organisms and the environment (Plutchik, 1980).

Emotions are difficult to model and capture because they change quickly. In this direction, exist three predominant approaches (*i.e.,* physiological, neurological and cognitive) that explain how emotions are generated. The first is about *physiological theories* that suggest changes in the body of humans generate emotions. In *neurological theories*, the hypothesis is that neuronal activity of the brain generates emotional responses. In contrast, *cognitive theories* describe that thought and mental processes generate emotional states in people (Lang and Bradley, 2010). Based on these approaches related to emotions, authors refer to three recognized models that classify emotions in humans. One of the most well-known works on the theory of emotions is the study of Ekman (1999). He tries to classify emotions into six discrete and primary categories (happiness, anger, fear, sadness, disgust, and surprise) and defines them as

universal and biologically basic.

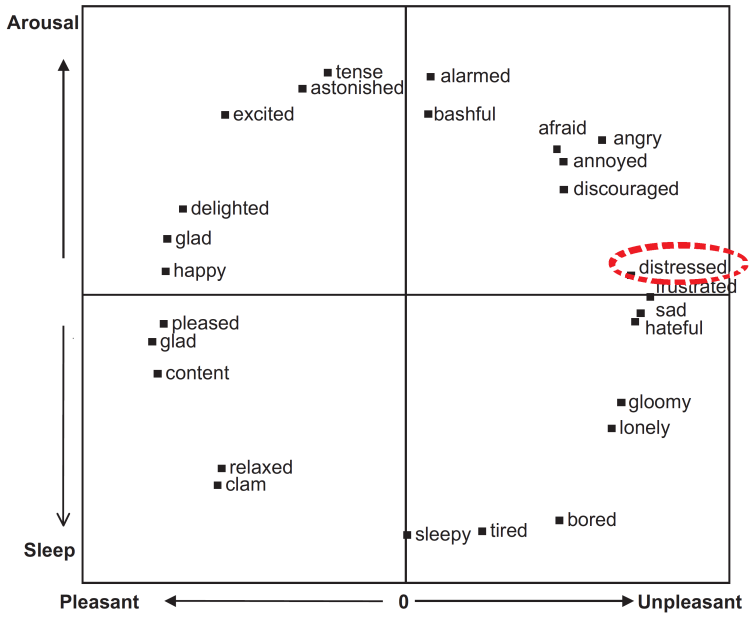


Figure 2.1: The circumflex model of affect, image retrieved from (Russell, 1980).

On the other hand, Russell (1980) proposed a two-dimensional circumplex model, which has the objective of categorizing emotions in continuous scales and basic dimensions, associated with arousal and valence (pleasant). This type of work is described in (Russell, 1980; Schlosberg, 1954). This model has been widely accepted in psychology field because it has been useful to identify emotions efficiently (Yoo et al., 2011). Figure 2.1 shows the circumplex model of affect, where the stress state is classified in the quadrant of *Arousal* and *Unpleasant* (red dashed lines). Stress was initially defined as an emergency response (Selye, 1936); for instance when human organism is in danger or injured, stress prepares the body and mind for the upcoming or current situation (stress reaction). Therefore, stress can generally be defined as a disruption of homeostasis whereby the anticipation of a stressful situation can also be enough to trigger a stress response (Ulrich-Lai and Herman, 2009; Herman et al., 2005).

2.2 Stress from electrodermal activity

Electrodermal Activity (EDA) is a psychophysiological parameter that reflects the activity of the sympathetic nervous system; also it could be interpreted as the level of activation of the subject. In other words, when the subject is very active (*i.e.,* high emotionality) the electrical conductance of the skin increases; on the contrary, when the subject is little activated (relax), the conductance of the skin decreases.

2.2.1 Anatomical and physiological basis

The skin is the largest and the heaviest organ of the human body. It provides very effective protection against the invasion of bacteria and other foreign substances. One of the principal functions of the skin is the maintenance of water balance, which corresponds to the regulation of body temperature. When you get warm, your skin releases sweat through sweat glands to reduce the temperature. Small blood vessels in your skin can also become filled with blood when you are cold, causing the temperature to increase.

There are two forms of sweat glands: *eccrine* and *apocrine*. Our work is focused on eccrine glands because they cover most of the body and are most present on palms and soles of the feet. According to Dawson et al. (2007), all eccrine glands are implicated in emotion-evoked sweating (usually most visible in areas with high gland density). An eccrine gland is composed of a compact coiled body which is the secretory portion and the sweat gland that is the long tube (the excretory portion of the gland). Figure 2.2 presents the general peripheral mechanism of EDA production[1].

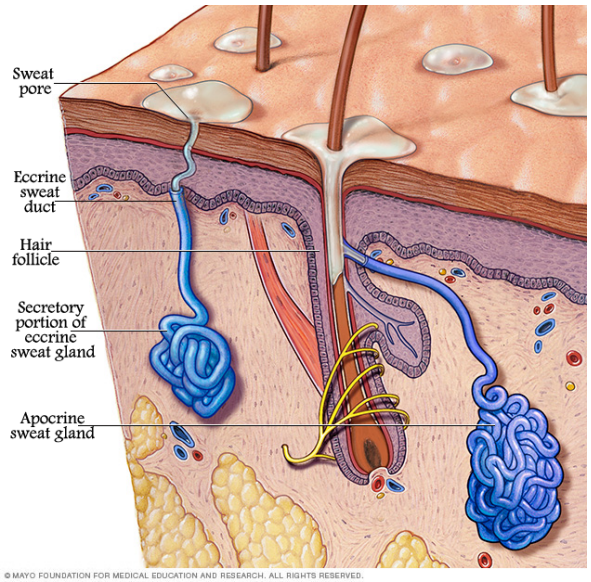


Figure 2.2: Anatomy of eccrine sweat gland. Image based on sweat glands from Mayo Foundation for Medical Education and Research[1].

Edelberg (1972), proposed a model to explain how peripheral mechanism is associated with the increase of skin conductance elicited by stimuli. He summarized that there are two peripheral mechanisms to help in the production of EDA: (1) production of sweat in the sweat glands and secretion of sweat by the sweat duct; and (2) activity of a selective membrane that is located in the epidermis. Moreover, autonomic nervous system (sympathetic and parasympathetic) is considered as potential mediators of EDA, this is due to the acetylcholine (parasympathetic neurotransmitter) that is

[1]Sweat glands: https://www.mayoclinic.org/diseases-conditions/hyperhidrosis/multimedia/sweat-glands/img-20007980

the mediator of eccrine sweat gland activity (Valenza and Scilingo, 2014). Finally, according to Boucsein (2012), thermoregulatory sweating associated with EDA is likely to hypothalamic activity; increase of muscle tone and gross locomotor skills associated with EDA are likely to activation of the reticular formation; affective processes associated with EDA are likely to amygdala activation; orienting attention associated with EDA is likely to prefrontal cortical activity; and fine motor skills associated with EDA are mediated by premotor cortex.

2.2.2 Recording system

EDA is measured by passing a small current through a pair of electrodes placed on the surface of the skin. The principle invoked in the measurement of skin resistance or conductance is the Ohm's law, which states that skin resistance (R) is equal to the voltage (V) applied between two electrodes placed on the skin surface divides by the current (I) being passed through the skin: that is, $R = V/I$. If the current is held constant, then one can measure the voltage between the electrodes, which will vary directly with skin *resistance*. Alternatively, if the voltage is held constant, then one can measure the current flow, which will vary directly with the reciprocal of skin resistance, skin *conductance*. Lykken (1971) argued strongly for the direct measurement of skin conductance with a constant voltage because skin conductance had been shown to be more linearly related to the number of active sweat glands and their rate of secretion. On the other hand, the overall resistance of a parallel circuit is a complex function of the individual resistances. Another recording issue concerns the hand from which to record. Many laboratories use the non-dominant hand for EDA measurements because it is less likely to have cuts or calluses and it leaves the dominant hand free to perform a manual task.

2.2.3 Data collection: E4-wristband

The E4-wristband[2] is a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition that provides biometrical information through its sensors: photoplethysmography, 3-axis accelerometer, optical thermometer and EDA (skin conductance). E4-wristband provides a way to collect data in an unobtrusive way in real-time (*e.g.,* collect data in the lab, home or in the wild). Also, this device has an application programming interface (API)[3] to develop our own app in real-time data-streams by Bluetooth connection based on Android. Figure 2.3 shows the technical specifications and location of E4-wristband sensors.

[2]E4-wristband website: https://www.empatica.com/en-eu/research/e4/
[3]E4-wristband for Developers: http://developer.empatica.com/

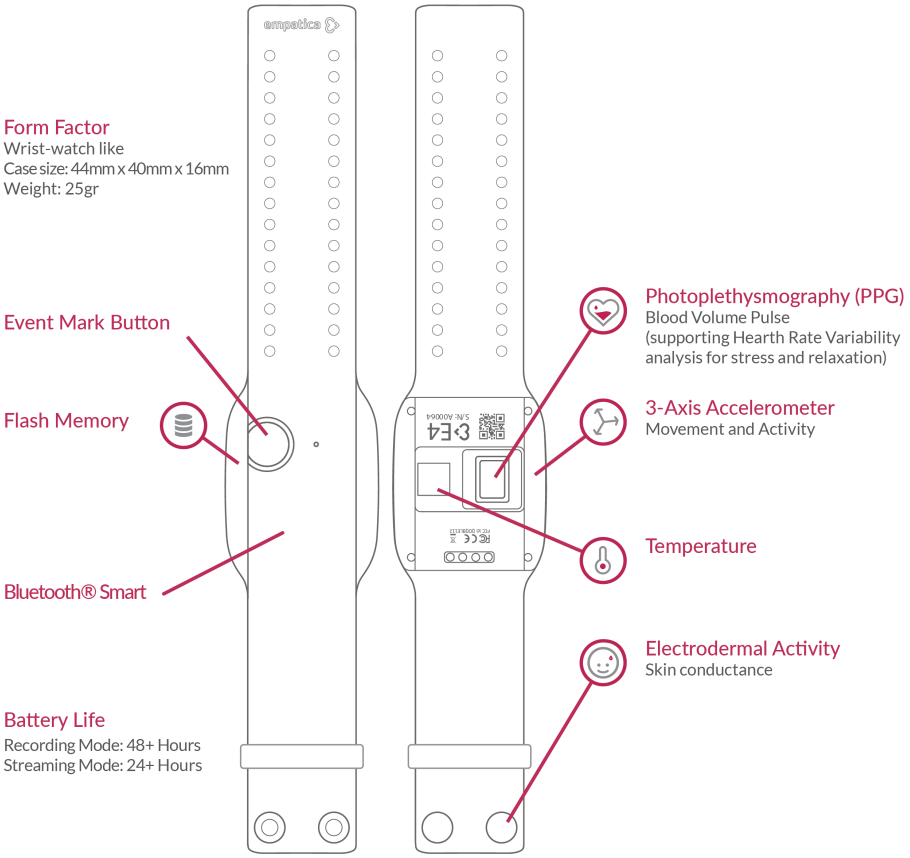


Figure 2.3: Technical specifications of E4-wristband[2].

2.3 Emotional triggers

In experimental settings, researchers can generate emotions on users intentionally, by using specific emotional triggers determined by the emotion to be induced. In this context, *an emotional trigger is any external or internal stimulus that elicits a negative or positive emotion (*e.g., *uncomfortable or comfortable temperature, environmental noise).* Nowadays, different kinds of emotional triggers exist; for instance, Kanjo et al. (2015) classified them into seven types: environment, physical movements, memories, perception, interacting with others, accomplishments and failure. Due to this research is focused on recognizing physiological stress as emotion target, stress triggers will be used to test our stress detector (first phase of results). These tests are performed using three types of stress triggers, the first is an environment trigger, where participants are exposed by five minutes to listen to fire alarm sounds (Westman JC, 1981; Passchier-Vermeer W, 2000). The second one is a social trigger (Sing-a-Song Stress Test), where participants are asked to sing a song aloud for 30 seconds with their arms still (Brouwer and Hogervorst, 2014). Finally, a cognitive trigger (Stroop Task), where participants have to pay attention and react to the color of a word while ignoring the word itself (Lattimore, 2001; Renaud and Blondin, 1997). A version of this trigger works with 4 colors; then the stimuli are the set of words: “green”, “red”, “yellow” and “blue” written in all four different colors. Additionally, the words are presented randomly for each

participant.

Chapter 3

Literature review

Nowadays, with the progress of new technologies and the introduction of interactive systems, the demand for software services that can understand and adapt their services according to user needs, have significantly increased. Also, with the purpose to understand whether those software services cause negative or positive emotions in their users, there have been developed many techniques to recognize human emotions using multiple data sources. In this direction, this Chapter presents a review of related works about user feedback and emotion recognition.

3.1 User feedback in software engineering

The assessment of user feedback is important for software evolution, and it concerns to involve old designs to generate news (Godfrey and German, 2008). In this context, the user feedback is interpreted as relevant information from any response about the interaction between the software system and the user (Madhavji et al., 2006). Besides, Morales-Ramirez et al. (2015) defined user feedback as "*a reaction of the user upon her experience in using a software service or application. It could be based on multimodal communication, such as natural language text, images or emoticons. Moreover, it contains meaningful information with the purpose of suggesting improvements (*i.e., *requesting new needs, reporting failures, and asking for modifications or clarifications)*".

During many years, user feedback (especially explicit) has been mainly used by recommendation systems (Morales-Ramirez et al., 2015), and they depend on any explicit feedback (*e.g.,* filling questionnaires, sending a report about an error, suggestion of new functionalities, or in the form of ratings) that users can provide. For that reason, explicit feedback is difficult to obtain (Hu et al., 2008). In this direction, many works have been proposed, such as (Zhao et al., 2018), (Maalej et al., 2016), (Lagun et al., 2013) or (Eisenstein and Puerta, 2000). For instance, Schefels and Zicari (2010) proposed a framework analysis for handling explicit feedback provided by guests of a website. Kim et al. (2016) collected explicit feedback on web search results from users

for analyzing whether a web search result is relevant.

3.1.1 Implicit user feedback

Researchers have proposed different approaches to deal with implicit feedback. Peska (2016) proposed a model of relevant contextual features for collecting implicit user feedback; as result of her research, a dataset of feedback of real e-commerce users was published. Lee et al. (2008) proposed to construct explicit feedback data (pseudo ratings) from implicit feedback data for an e-commerce recommender system. Leiva (2011) captured touch events (implicit feedback) for adapting, rearranging and restyling the interacting items of a web browser. Claypool et al. (2001) analyzed the correlation between implicit and explicit feedback. For that, they developed a web browser for collecting implicit (*i.e.,* elapsed time, mouse movement, mouse clicks and scrolling) and explicit (ratings) feedback. In the direction of mixed models that combines implicit and explicit feedback, can be found in some works such as (Li and Chen, 2016), (Zanker and Jessenitschnig, 2009), (Liu et al., 2010), or (Jawaheer et al., 2014). For instance, in the context of music recommender systems, Jawaheer et al. (2010) carried out some experiments for comparing explicit and implicit feedback datasets. Their results shown that implicit and explicit positive feedback have similar performances despite different characteristics (*e.g.,* accuracy, abundance, context-sensitive, expressivity of user preference), and conclude that both types of user feedback should complement each other. For facilitating this combination, some platforms have been implemented for gathering multimodal feedback (*e.g.,* Seyff et al., 2014; Schneider, 2011). Additional examples can be found in the existing literature of implicit feedback (*e.g.,* Jaramillo Garcia et al., 2015; Vrochidis et al., 2011; Song and He, 2010; Hu et al., 2008).

3.2 Emotion recognition

During many years, researchers in several computer science fields have paid attention to develop methods for recognizing and understanding human emotions. For instance, based on natural language processing (Liu et al., 2013; Tang, 2015; You, 2016), emotion recognition through facial expression (Busso et al., 2004; Le and Vea, 2016; Menne and Lugrin, 2017) or using physiological data (Gouizi et al., 2014; Girardi et al., 2017; Canento et al., 2011; Guendil et al., 2015; Healey and Picard, 2005; Mozos et al., 2017; Sriramprakash et al., 2017; Garcia-Ceja et al., 2016; Sano and Picard, 2013; Bogomolov et al., 2014).

There exist some research works focused on stress detection using only physiological data. For instance, Mozos et al. (2017) proposed to combine machine learning techniques using EDA, Photoplethysmogram (PPG) and Heart Rate Variability (HRV) signals to detect stress in social situations using The Trier Social Stress Test (TSST) as stressful. Garcia-Ceja et al. (2016) used Accelerometer (ACC) data of a mobile phone to recognize stress in real workplace environments of thirteen subjects using two

classification models: naive Bayes and decision trees. They obtained an accuracy of 71%, and their study lasted 8 weeks.

Sano and Picard (2013) implemented different machine learning classifiers to detect stress: Support Vector Machine (SVM) with linear kernel, SVM with Radial Basis Function (RBF) kernel, k-nearest neighbors, Principal component Analysis (PCA) and SVM with RBF kernel and k-nearest neighbors. Their work is focused on comparing the performance of the implemented algorithms using the collected data of the subjects (Skin Conductance (SC), ACC, and mobile phone usage) of five days. Kocielnik et al. (2013) described a framework to detect stress in the context of a person's activities. They use a min-max algorithm and ACC as source data. Bogomolov et al. (2014) collected mobile phone activity (*i.e.,* call log, SMS log, Bluetooth interactions) of 117 subjects to recognize stress during common daily activities. They applied different classifiers: SVM, artificial neural networks, an ensemble of tree classifiers based on a Breiman's Random Forest (RF) and Friedmans Generalized Boosted Model (GBM). Similarly, using a range of machine learning techniques, some other examples can be found in the existing literature (*e.g.,* Sriramprakash et al., 2017; Sandulescu et al., 2015; Bauer and Lukowicz, 2012; Carneiro et al., 2012; Healey and Picard, 2005; Sun et al., 2012).

Table 3.1: Comparative chart of the most representative related works in stress recognition.

Author	**Classification algorithm**	**Source data**	**Evaluation tools**	**Context**
Mozos et al. (2017)	SVM, AdaBoost, and k-nearest neighbor.	EDA, PPG and HRV.	Accuracy of 89.75%, precision of 89.5% and recall of 95%.	Social situations using the TSST.
Garcia-Ceja et al. (2016)	Naive Bayes and decision trees.	ACC.	Accuracy of 71%.	Real working environments.
Sano and Picard (2013)	SVM, RBF, k-nearest neighbors and PCA.	SC, ACC and mobile phone usage.	Accuracy of 75%.	Stress detection that subjects are able to perceive and report.
Kocielnik et al. (2013)	Min-max algorithm.	SC and ACC.	No reported	Subject's activities.
Bogomolov et al. (2014)	SVM, ANNs, tree classifiers based on RF and GBM.	Mobile phone activity	Accuracy of 72.39%.	Common daily activities.

3.3 Final considerations

Table 3.1 summarizes the most representative related works, illustrating the diversity of used algorithms to recognize stress. Most of these works use a machine learning method to implement the classifier; however, there can be some issues of using these methods: there exist a lack of benchmark datasets of physiological data needed for training the model and for validating results. Also, these methods need big datasets to carry out the training stage, where the machine learns about the user behavior in relationship with predefined tasks. To address this issue regarding the training of stress detector, this research uses an arousal-based statistical approach for detecting stress in real-time. This statistical algorithm introduces two main advantages for the resulting stress detector: i) reliability is independent of a training dataset, in contrast to the requirement imposed by approaches based on machine learning algorithms; ii) higher flexibility is provided since the detector can be used in different user conditions. The following Chapter presents the algorithms used for detecting physiological stress and for processing the user context.

Chapter 4

Actionable emotion detector

This thesis contributes to the goal of the HAPPYNESS framework (Condori-Fernandez, 2017), through the implementation of its *emotion measurement* module. HAPPYNESS proposes to exploit the emotional information of users to provide a personalized context-aware software service with the objective to enhance the quality of User Experience (UX). This chapter presents the used algorithms for detecting actionable emotions. In Section 4.1, the requirements of the actionable emotion detector are investigated. Section 4.2 presents the software architecture of the proposal. The method for detecting physiological stress is presented in Section 4.3. Section 4.4 shows the procedures to process the user context. Finally, the base rules of the inference engine for determining whether detected stress is actionable are presented in Section 4.5.

4.1 Requirements of the actionable emotion detector

According to the HAPPYNESS framework (Condori-Fernandez, 2017), the actionable emotion detector is challenging to two quality aspects of context (*i.e.,* freshness and temporal resolution). Based on the challenges proposed by Condori-Fernandez (2017), five requirements were identified (*i.e.,* accuracy, scalability, flexibility, context completeness and freshness), which are related to each other. Overall, these five requirements were addressed for the implementation of the actionable emotion detector, and are explained in the following sections. Figure 4.1 shows the dependency relationship between the requirements (Re) to ensure the correct working of the actionable emotion detector.

Nowadays, it is possible to detect different types of human emotions from different data sources (*e.g.,* EDA, HRV, Blood Volume Pulse (BVP), Electrocardiogram (EKG)). In the same context, the user environment can be modeled in different scenarios according to the availability of sensors. In this direction, the requirements presented in Figure 4.1 are defined as:

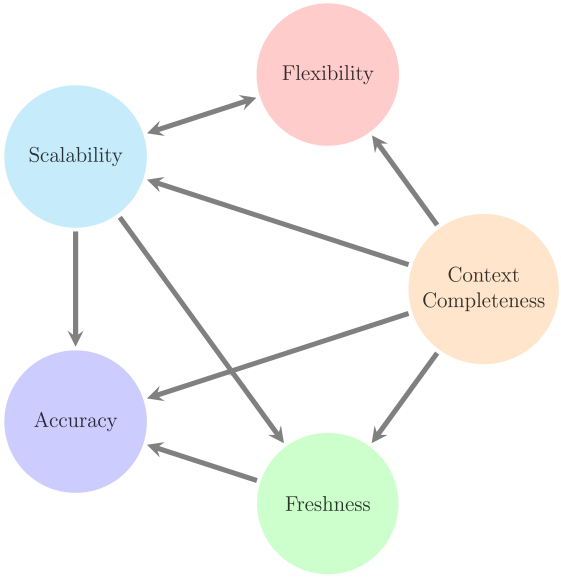


Figure 4.1: Relationship between requirements of the actionable emotion detector.

- **Re1. Flexibility:** the architecture of the actionable emotion should be flexible to add different emotional and environmental sensors to recognize different emotions and to model the user context.
- **Re2. Scalability:** this requirement is related to the previous requirement; while different sensors are added, the actionable emotion detector should work correctly. However, adding more sensors means actionable emotion detector will process more data that could affect directly to the *freshness* and *accuracy* requirements. Another issue about this requirement is to know what extent it can be scalable? That depends on many factors, such as memory or connectivity, being the main issue the hardware architecture of the devices.
- **Re3. Accuracy:** it refers to how accurately the real-time emotion detector can recognize emotions in subjects, but also it is related to the precision of filtering external threats (emotional triggers) for the assessment of the actionable emotion. This requirement is influenced by the scalability, freshness, and context completeness requirements.
- **Re4. Freshness:** this requirement refers to the time that elapses between the software service occurred and the moment when the actionable emotion detector recognized it. This interval should be short to achieve a detection in real time. Consequently, the method used to detect user emotions should help to accomplish the detection in real-time (see section 4.3). Whether this time is long, the accuracy will be low.
- **Re5. Context completeness:** in the scenario for recognizing emotions in a common interaction between the user and software service, the context of the user could influence his emotion perception. For that reason, this requirement is associated with the three before mentioned. It indicates the actionable emotion

detector should have the capability to model emotional threats of the user context (see Section 4.4).

4.2 Architecture

Based on the requirements before defined, Figure 4.2 shows the proposed architecture for detecting actionable emotions, which is mainly integrated by the following components:

4.2.1 Context-aware system

The software system provides the context-aware service to improve the interaction with the user. Moreover, it notifies the inference engine the moments when the service is emitted to enable the assessment of the actionable emotion. On the other hand, the context-aware service could be considered as an emotional trigger because it depends on the service configuration, and could generate on user a negative/positive emotion. However, the possible emotions generated by the service are considered as useful information (user feedback) for all stakeholders.

4.2.2 Internal context

This module works together with the *real-time emotion detector* component of the actionable emotion detector module; they have as functions the collection and processing of signals (from the sensors) to recognize user emotions. In other words, the principal task for this module is to detect emotion states. Furthermore, the sensor component is not limited to only one sensor; it could be integrated with many sensors that are inputs to the preprocessing unit. In this thesis, physiological stress was selected as the emotion to the assessment. For that reason, skin conductance signals from the EDA sensor of the E4-wristband are used. Section 4.3 gives more details about the applied methods in this task.

4.2.3 External Context

Both the external context module and the *context analyzer* have the function of collecting and processing information from the external context of the user to help in the assessment of the actionable emotion. The context can be modeled by n emotional triggers; likewise, an emotional trigger can be modeled by n sensors.

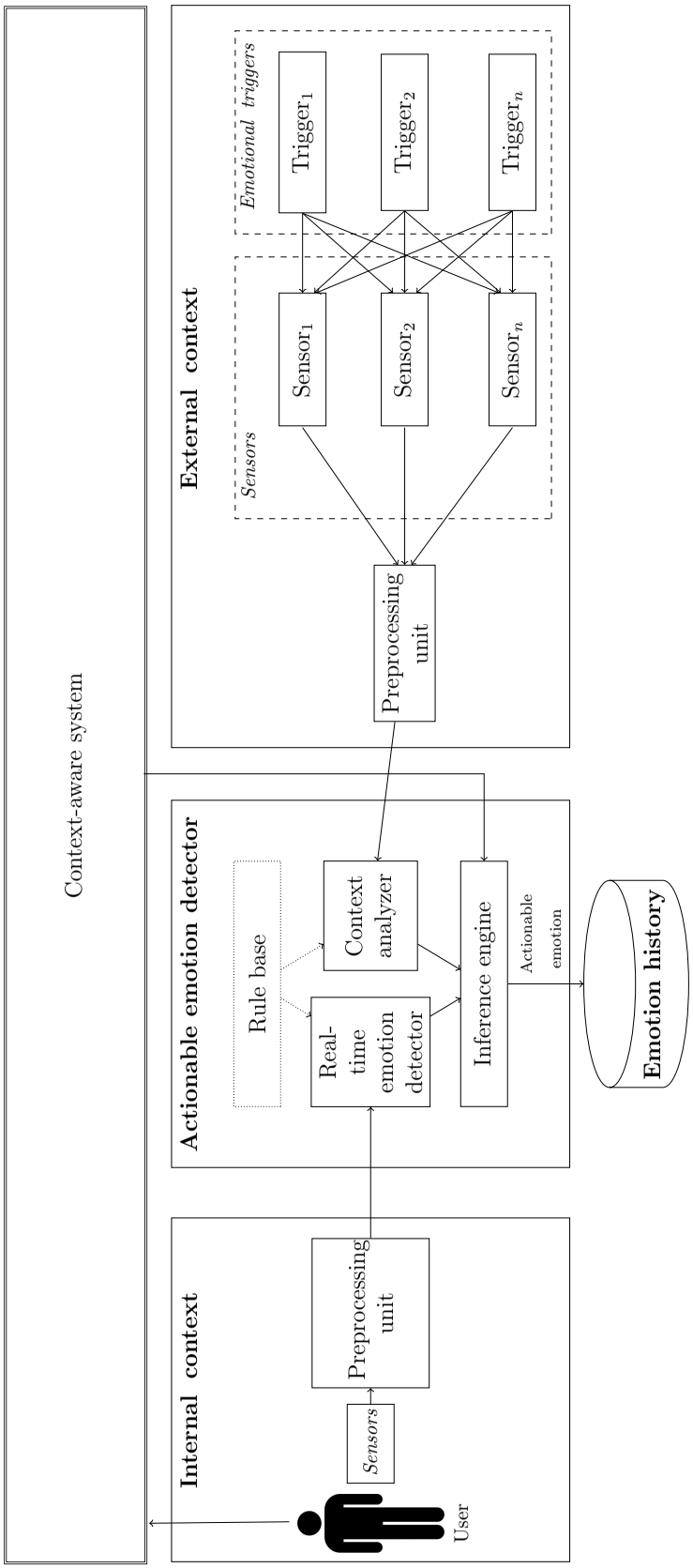


Figure 4.2: The software architecture of the actionable emotion detector.

However, the number of sensors supported by the actionable emotion detector depends on the scalability requirement (highly related to the hardware architecture). For this thesis, the external context of the user is modeled to recognize three emotional triggers: i) *environmental noise* that uses the microphone of the mobile phone; ii) *interaction with others* using the camera and microphone, and iii) *physical movements* that uses the camera of the mobile phone and the accelerometer sensor from the E4-wristband. Section 4.4 describes the methods used to process the output of each sensor.

4.2.4 Actionable emotion detector

The inference engine has as inputs the processed data by the real-time emotion detector, the context analyzer components (previously defined), and the context-aware system. Based on the predefined rules, this module determines if the detected emotion is actionable or not. Section 4.5 presents the rules of the assessment for the actionable emotion.

4.2.5 Emotion history

Finally, the recognized actionable emotions are saved in an emotion historic database, with the objective to give valuable information to HAPPYNESS for learning about the user emotions, and it can set the emotional thresholds. The emotion history has not been implemented because it is not part of the scope of this thesis.

4.3 Real-time stress detector

This section addresses the accuracy and freshness requirements (**Re3** and **Re4**), that is related to accomplish an emotion detection in real-time. Figure 4.3 shows the stress detection process of the approach that has been automated to detect stress of individuals (*e.g.,* programmers, testers) in real-time. It is used wearable sensors (*i.e.,* E4-wristband[1]) to collect physiological data. The method is focused only on sensing electrodermal activity (EDA) as a main input for the implementation of the stress detector. A transient increase on the EDA signal is proportional to sweat secretion and it is related to stress (Bakker et al., 2011; Boucsein, 2012).

The main functionality of the stress detector is to determine whether the user is stressed or not. The detector will mark a label of "*stressed*" or "*not stressed*". We have implemented the preprocessing steps proposed by Bakker et al. (2011) for arousal detection in an integrated pipeline to enable real-time processing (see Figure 4.3 for the involved preprocessing steps). Next, we explain the methods/algorithms that were used in the stress detection process.

[1]E4-wristband website: https://www.empatica.com/e4-wristband

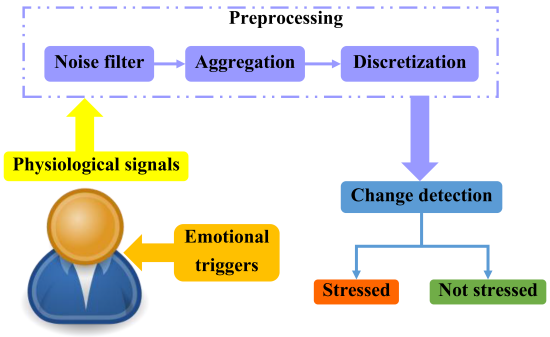
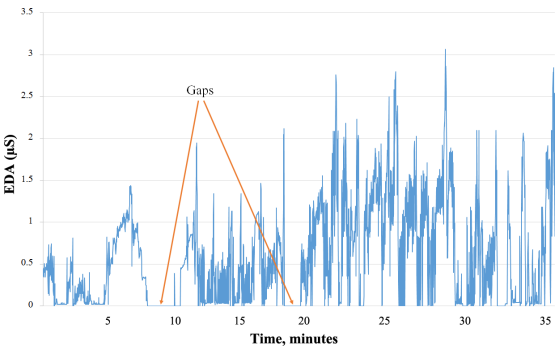


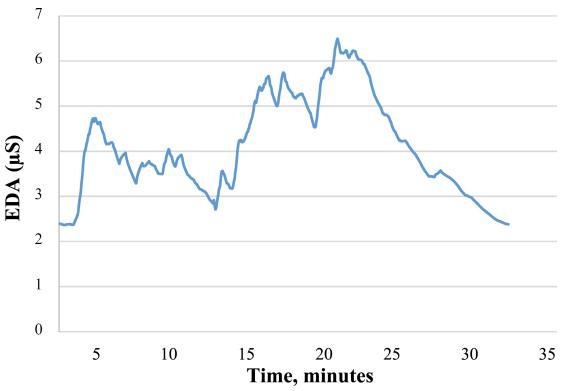
Figure 4.3: Overview of the stress detection process.

4.3.1 Noise filter

To recognize stress changes is used EDA signals, then the first stage in the pipeline of the stress detector is to collect raw signals by the Empatica's E4-wristband, Figure 4.4a presents a common sample of EDA signals, which is measured in microSiemens (μS), a unit of electric conductance. Usually for measuring EDA is required two electrodes that need skin contact to produce a reliable signal, therefore the quality of the collected EDA signals depends on the continuity of the contact between user skin and the device's sensors. However, the contact is not the same in all users and noise could be introduced in the signal. Hence, noise filtering is needed to mitigate these issues in the input (*e.g.*, in Figure 4.4a, we can find some gaps as a consequence of weak skin contact). Before analyzing EDA signals, it is important to clean raw data, because noise might be mistaken as genuine peaks. Therefore, the first step of the preprocessing is to apply a median filter over a moving window of size $n = 100$ EDA samples, as suggested in (Bakker et al., 2011). Figure 4.4b shows the noise filtering of the collected raw data.



(a) Raw EDA signals.

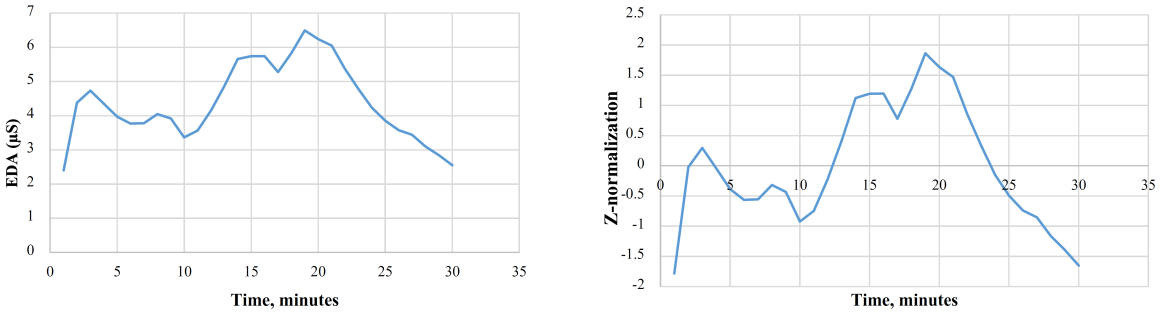


(b) Filtered EDA signals.

Figure 4.4: Processing of EDA signals to remove noise. (a) Gaps occur when the contact between the user skin and the sensors is not tight. (b) Clean raw EDA signals after applying a noise filter.

4.3.2 Aggregation

The EDA signal acquired by the E4-wristband is sampled at 4Hz (*i.e.,* the device provides 4 samples or readings per second, which means 240 samples per minute).



(a) Aggregation. (b) Normalization.

Figure 4.5: Aggregation and normalization of EDA signals. (a) Aggregated process over previous filtered data. (b) Z-normalization of aggregated data.

Based on (Bakker et al., 2011), it is applied an aggregation step of each minute over the filtered input signal: given y' is a moving window of size $m = 240$ (the EDA samples of one minute), where $y_1, ..., y_m$ is aggregated to a single value y'' where $y'' = max(y')$. For instance, Figure 4.5a shows the aggregation of approximately 7200 filtered EDA samples to 30 representative points (collected signals of 30 minutes).

4.3.3 Discretization

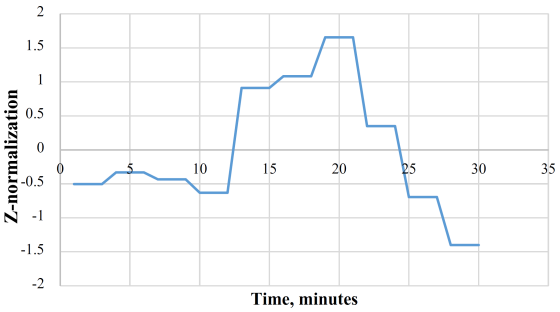
In this step, the data is discretized using the symbolic aggregate approximation (SAX) method (Lin et al., 2003). It is a means for very efficient local discretization of time series subsequence from 1 to 5 that can be interpreted as levels of stress variation (1: completely relaxed to 5: maximum arousal). Those levels should not be understood as absolute levels of arousal, but rather as a local relative measure of arousal.

The input for SAX is a time-series X of length n and the output is a string of length w, where $w < n$ typically, the output string is normalized to an alphabet of size > 2. The algorithm consists of the following two stages:

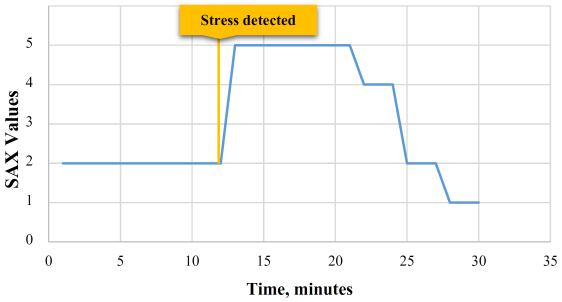
- Transformation of original time-series into a Piecewise Aggregate Approximation (PAA) representation. To do this, first it is necessary a Z-normalization (see Equation 4.1), where the mean is around 0 and the standard deviation is close to 1, using the following formula:

$$x'_i = \frac{x_i - \mu}{\rho} \qquad (4.1) \qquad\qquad \overline{\mathrm{x}}_i = \frac{M}{n} \sum_{j=\frac{n}{M}(i-1)+1}^{(\frac{n}{M})i} x_j \qquad (4.2)$$

Where μ is the mean of the time series and ρ is the standard deviation. After the Z-normalization, we can apply PPA transform, which approximates the time-series into vector $\overline{X} = (\overline{x}_1, ..., \overline{x}_M)$ of length $M \leq n$ (see Figure 4.5b). Where each $\overline{x}_i$ is calculated with the Equation 4.2. With the objective to reduce the dimensionality from n to M, first we divide the time-series to n/M equally sized samples and calculate the mean for each sample (see Figure 4.6a).



(a) PAA tranformation.



(b) SAX transformation.

Figure 4.6: PAA and SAX transformations of EDA signals. (a) PAA representation of the preprocessing data. (b) *SAX* representation and stress detection using ADWIN algorithm.

- Transformation of the PAA data into a string. The method use a breakpoint or cuts $B = \beta_1, \beta_2, ..., \beta_{\alpha-1}$ such that $\beta_{i-1} < \beta_i$ and $\beta_0 = -\infty, \beta_\alpha = \infty$ divides the total area in equal subareas. Additionally, it assigns a symbol $alpha_j$ to each interval $[\beta_{j-1}, \beta_j)$, and the final conversion from PAA coefficients $\bar{C}$ into a SAX string $\hat{C}$ is with the Equation 4.3. Figure 4.6b shows SAX transformation of the previous preprocessing signals.

$$\hat{c} * i = alpha * j, iif, \bar{c} * i \in [\beta_{j-1}, \beta_j) \tag{4.3}$$

4.3.4 Change detection

We use a change detection algorithm based on ADaptive WINdowing (ADWIN) method (Bifet and Gavaldà, 2007). ADWIN computes the mean for each split of a sequence of signals and analyzes the statistically significant difference between two consecutive splits. When a statistically significant difference is detected at point p_i, ADWIN drops the data backwards from p_i, after it repeats the splitting procedure until no significant differences be found in the sequence. For instance, given ϕ_1 and ϕ_2 as the means of two splits of a sequence of EDA signals, then $|\phi_1 - \phi_2| > \epsilon_{cut}$ is the condition for a change detection that is computed with the Equation 4.4.

$$\epsilon_{cut} = \sqrt{\frac{2}{m}.\sigma_W^2.ln\frac{2}{\delta'}} + \frac{2}{3m}ln\frac{2}{\delta'} \tag{4.4}$$

where σ_W^2 is the variance of the elements of W. δ is the desired confidence and

$\delta' = \delta/(ln\ n)$ (Bakker et al., 2011). Figure 4.6b shows the output the algorithm detecting a stress change.

4.4 Context analyzer

Modeling the user context is a complex task that requires to process a massive amount of data from different sensors of the user environment. Also, different agents (*i.e.,* emotional triggers) from the user context could generate negative emotions in the user (*e.g.,* stress). Therefore, stress episodes generated by these emotional triggers need to be recognized and excluded from the assessment of the actionable emotion detector. For providing reliable user feedback, three emotional triggers for modeling the user context were selected: environmental noise, interaction with others and physical movements (selected from the classification presented in (Kanjo et al., 2015)). Since the context analyzer addresses the context completeness requirement (**Re5**), and it is directly related to flexibility (**Re1**) and scalability (**Re2**) requirements, the actionable emotion detector performance does not have to be affected while more sensors are added to filter threats of the user context. The following subsections present how the three emotional triggers have been modeled.

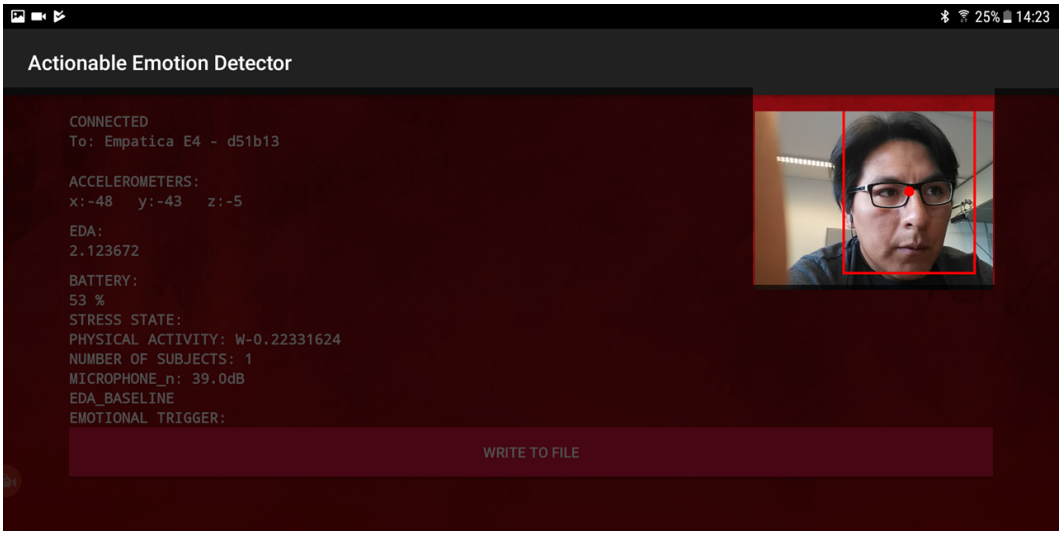


Figure 4.7: An example of the adapted face recognition algorithm.

4.4.1 Sensing and analyzing environmental noise

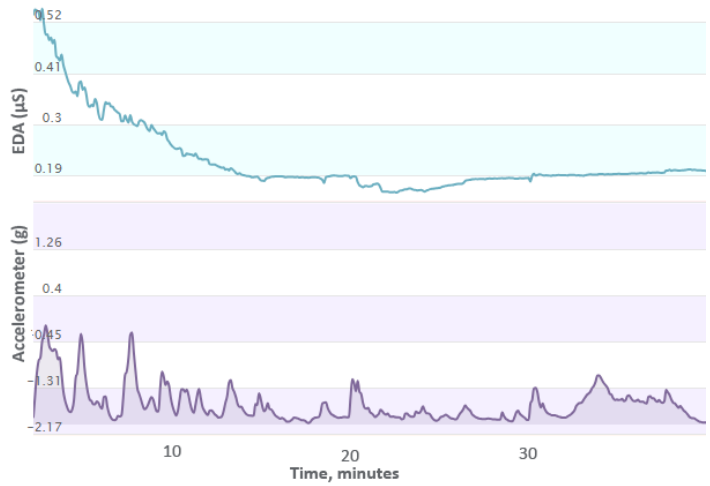
There are many emotional triggers associated to the environment (*e.g.,* temperature, sounds, lighting). One of the mostly studied causes of stress and often used as stressor is the environmental noise (*e.g.,* Walker et al., 2016; Morrison et al., 2003; Passchier-Vermeer W, 2000; Lercher et al., 1993; Westman JC, 1981). To this purpose, the smartphone's microphone sensor is used to gauge the environmental noise. Captured

sound is transformed into a decibels (dB) scale; then thresholding is applied to classify the sound [2]; *i.e.,* a sound is considered noise when $> 80dB$ (Davis and Davis, 1997).

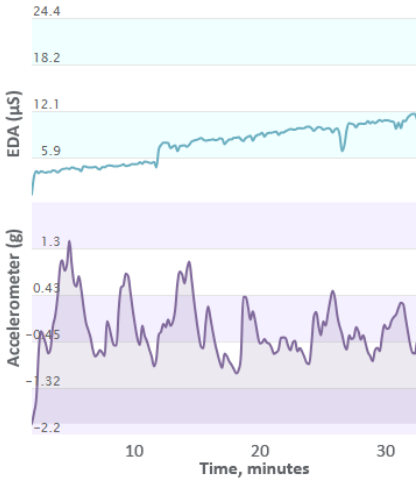
4.4.2 Sensing and analyzing the social interaction with others

We restrict the emotional trigger to interactions with other people face to face in working places, leaving apart possible remote interactions. Different studies (*e.g.,* Thanh, 2016; Page-Gould et al., 2010; Azari et al., 2010; Hartley and MacLean, 2009) indicate that this kind of interactions could generate stress episodes depending if the other person gives an adverse notice, or it could generate other stressful situations (*e.g.,* an extra office work).

To detect other people interacting with the user, our work uses both the camera and microphone sensors in combination. In particular, we used a face recognition algorithm based on the *GraphicFaceTracker* class provided by Android[3] to track the number of users involved, using the front camera to acquire the camera frame. Then, the user is considered to be interacting with other people when more than one person are detected and decibels level are between 40dB and 80dB (an ordinary conversation). Figure 4.7 shows a testing sample screen with the decibel classification from the microphone and one user detected in the current frame.



(a) Low EDA values (0.19μS) with a low physical activity (sitting).



(b) High EDA values (9.5μS) with a high physical activity (running).

Figure 4.8: Relationship between EDA signals and physical activity.

[2] http://www.aes.org/
[3] https://developers.google.com/vision/android/face-tracker-tutorial

4.4.3 Sensing and analyzing physical movements data

The physical movements in the office are integral parts of daily working activities. It is well-known that physical activity activates physiology and can easily confound the assessment of stress (Plarre et al., 2011). Different works report that the reliability of stress assessment reduces significantly in the presence of physical activity, even in a lab environment (*e.g.,* Rahman et al., 2014; Bakker et al., 2012; Hong et al., 2012; Plarre et al., 2011). Figure 4.8 shows an example of how EDA values (light blue signals) increase when physical activity increases (purple signals).

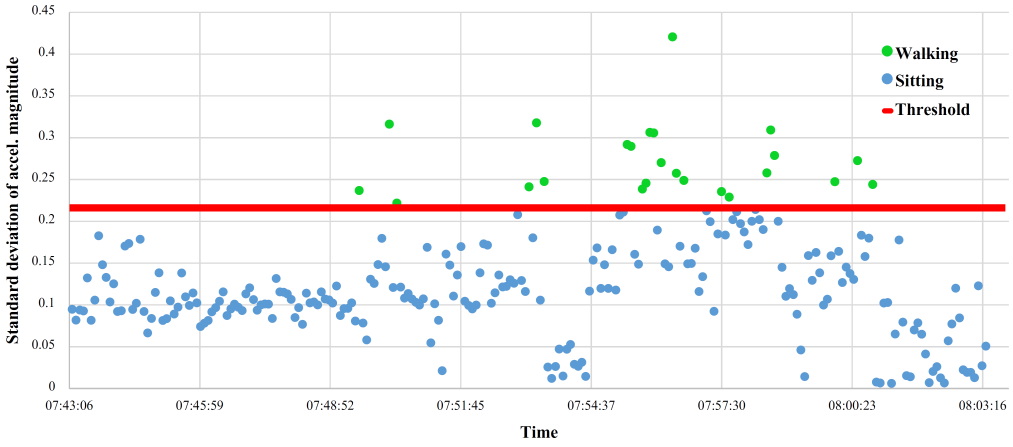


Figure 4.9: Classification of physical activity of an anonymous user.

For this objective, we use the accelerometer sensor of the E4-wristband and we adopted the proposal developed by Atallah et al. (2010) for detecting physical activity changes. This method uses the data of the accelerometer sensor for classifying the physical activity using a threshold. According to accelerometer sensor specifications, E4-wristband has an onboard MEMS type 3-axis accelerometer that measures constant gravitational force (g), applied to each of the three spatial dimensions (x, y, and z), in the range $[-2g, 2g]$. Following the preprocessing steps of Empatica[4] for filtering the raw data of the acceleromenter sensor, it is needed to normalize the spatial dimensions: $x/64$, $y/64$ and $z/64$.The processing of accelerometer data includes a median filter over a moving window of size = 100 and drift removal. After that, we compute the magnitude of the three spatial dimensions for each sample using the following Equation:

$$m = \sqrt{x^2 + y^2 + z^2}$$

Next, we compute the standard deviation of magnitude in windows of size $N =$ 1920 logs (60 segs):

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

[4]The technical specifications and how to process the raw data of the accelerometer sensor is detailed in https://support.empatica.com/hc/en-us/articles/202028739-How-is-the-acceleration-data-formatted-in-Empatica-Connect-

Then, if the standard deviation of accelerometers magnitude is greater than 0.21348 (this threshold is defined by Atallah et al. (2010)), this window is labelled as moving (*i.e.,* walking or running) and the others are labelled as static. Figure 4.9 shows the classification of the physical activity using a threshold.

Finally, to help with the assessment of the physical movements, the information of the accelerometer sensor is combined with the knowledge of the number of users provided by the adapted face recognition algorithm. In this context, it is considered as moving state only when the accelerometer evaluation indicates a moving state and camera registers that there is no exist any person in the current scene (Doherty et al., 2013).

Table 4.1: Possible combinations of internal context and external context variables for the assessment of the actionable emotion.

		Emotional trigger (stressor)				
	Software service	**Environmental noise**	**Interacting with others**	**Physical movements**	**Stress detector**	**Result**
Case$_1$	Delivered	Non-Detected	Non-Detected	Non-Detected	True	Actionable emotion
Case$_2$	Delivered	One or more triggers detected			True	Undetermined
Case$_3$	Non-Delivered	Any trigger detected			True	Emotion detected
Case$_4$	Non-Delivered	Non-Detected	Non-Detected	Non-Detected	False	Non-emotion

4.5 Inference engine and rule base

The inference engine is continuously receiving information from the real-time emotion detector (stress detector, in this case) and the context analyzer; with these inputs, the inference engine has three primary rules for the assessment of the actionable emotion:

- **Case$_1$**, when the software service is delivered **AND** the emotion detector detects an emotion (TRUE) **AND** the context analyzer does not detect any emotional trigger **THEN** the emotion is actionable.

- **Case$_2$**, when the context analyzer detects one emotional triggers **AND** the emotion detector detects an emotion **AND** the software service was delivered **THEN** the result is an undetermined status.

- **Case$_3$**, when the context analyzer detects one emotional triggers **AND** the emotion detector detects an emotion **AND** the software service was not delivered **THEN** the result is an emotion detected (non-actionable).

The $case_2$ is defined as undetermined because with the obtained information the cause of stress is not possible to determine. The $case_4$ gives as result a non-emotion when the emotion detector does not recognize any emotion. Table 4.1 summarizes the four possible cases for the assessment of the actionable emotion.

4.6 Final Considerations

The implementation of the actionable emotion detector was divided into two stages, and it was programmed in Android[5], because for managing the sensors of the E4-wristband is required to install the Empatica API, which needs an Android mobile for connecting with the E4-wristband. To achieve the connectivity between a mobile phone and the E4-wristband, the application needs to initialize the EmpaLink library with an API key. In the first stage, only was implemented the real-time detector for the experiment explained in Chapter 5 and it was deployed in a Samsung J5 with Android 5.1-Lollipop. For the second stage, the context analyzer was added, and it was deployed in Samsung Tables for the experiment, which is explained in Chapter 6

This Chapter presented the software architecture for detecting actionable emotions and was defined that the assessment of the actionable emotion is based on physiological stress. Also, the three emotional triggers (environmental noise, interaction with others and physical movements) selected to evaluate the user context were explained in detail.

In one hand, the real-time stress detector mainly addressed the accuracy and freshness requirements; the human emotion detection using physiological is reliable but complicate because each person reacts differently. In this direction, finding an acceptable time for the freshness requirement without affecting the accuracy requirement is a difficult task. After many experiments with the stress detector method, the freshness was sacrificed (to three minutes) to get a balanced accuracy (that will be explained in the following chapter). On the other hand, the context analyzer addressed the flexibility, scalability, and context completeness requirements; the scalability is limited to the capability of the mobile devices. It showed that use of the camera, microphone, sensors, Bluetooth and internet connections at the same time was the maximum point of scalability that mobiles architectures can support.

[5]The source code can be found at https://goo.gl/q1iUZy

Chapter 5

Evaluation of the stress detector

This Chapter presents the design and results of the experiment to evaluate only the real-time stress detector. The stress detector was implemented in a single pipeline suitable for real-time processing following an arousal-based statistical approach. It works with physiological data gathered by the E4-wristband, which registers Electrodermal Activity (EDA). This experiment was conducted to analyze the output of the stress detector with regard to the self-reported stress in similar conditions to a quiet office workplace environment when users are exposed to different emotional triggers. This Chapter describes the experiment, next to the procedure of the experiment, then the threats to validity and finally the results.

5.1 Design of the experiment

This experiment was designed to validate the implemented stress detector, where participants experienced different stressful situations caused by emotional triggers. The goal of the experiment is to evaluate the performance of the stress detector in terms of its accuracy. This evaluation was performed from the viewpoint of office workers in the context of performing certain tasks that cause stress (emotion trigger). From this goal, the following research question is derived:

RQ_1: *How accurately is the stress detector able to recognize subjects stress under different types of emotional triggers?*

Based on the defined research question, **independent variables** *emotional trigger* were identified, originally with 3 levels (environmental: fire alarm; cognitive: Stroop Task; and social: Sing-a-Song Stress Test). After running a pilot study, the social trigger was removed to reduce the length of the experiment to thirty minutes. This is further explained in the data collection section. Figure 5.1 (c) shows a screenshot of the instructions for the Stroop Task.

As **dependent variables:** *subject stress status*, which is measured in a nominal

scale (stressed or not stressed); and *perceived stress* measured by means of a self-response questionnaire.

The hypothesis is that *when different types of emotional triggers are delivered, the stress detector is able to recognize stress with a similar accuracy.* Accuracy refers to the closeness of a measured value to a "true value". In this study, the true value of perceived stress was determined by the subjects of the experiment.

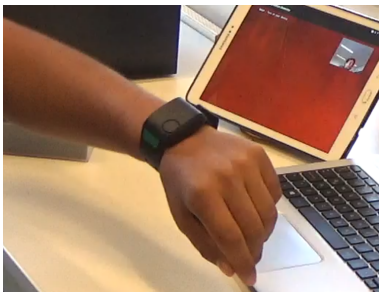
5.2 Subjects

Twelve subjects from University of Twente (The Netherlands), involved in research in computing areas (*i.e.,* Master students, PhD candidates), participated voluntarily in the experiment, whose ages ranged between 21 and 32 years old. Seven are women and five men.

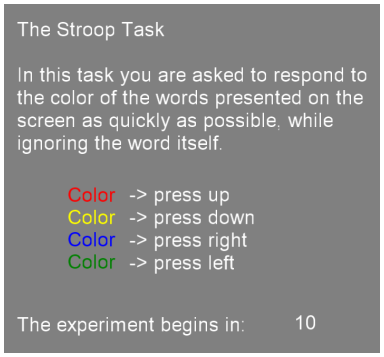
5.3 Instrumentation and procedure

The experiment was carried out in a quiet room equipped with a table and a chair as shown in Figure 5.1a. Subjects interacted with a laptop where the Stroop Task (cognitive trigger implemented with Psychopy[1]) was installed. Also, subjects wore the E4-wristband and headphones to interact with the environmental trigger. Figure 5.1b shows the correct position of the E4-wristband on the non-dominant hand of the subject.

(a)



(b)



(c)

Figure 5.1: Requirements to carry out this experiment. (a) A subject in the experiment room interacting with emotional triggers. (b) E4-wristband placed on the non-dominant hand. (c) Instructions for interacting with cognitive trigger (stroop task).

[1] http://www.psychopy.org/

The evaluation followed a within-subjects design, where all subjects were exposed individually to both cognitive and environmental triggers (treatments). The order in which the subjects interacted with the treatments were assigned randomly. Figure 5.2 shows the procedure of the experiment that consists of two phases:

Phase 1. Firstly, the subjects were asked to read and sign the informed consent form, which described the purpose and structure of the experiment. Subjects were informed beforehand about the sensing device and the possibility of experiencing some stress during the experiment. Furthermore, they were informed that they could pass on the task at any time if they considered stress unbearable. After signing the consent form, each subject got put on and adjusted the E4-Wristband to enable the gathering of physiological data. Then subjects were asked to complete a demographic questionnaire, and press a button to start the experimental tasks when they were ready. This phase lasts around five minutes.

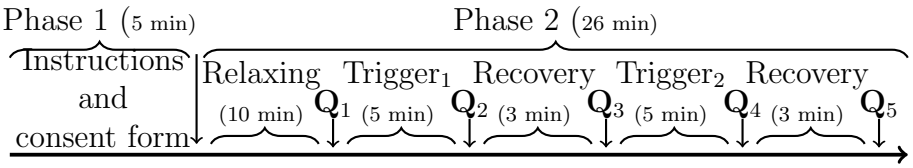


Figure 5.2: Experiment procedure and timeline.

Phase 2. Each subject was asked to sit on her/his own chair in a comfortable position for 10 minutes. During this period, they were asked to stay quiet and relaxed. Then subjects interacted with the corresponding treatments (five minutes each). Also, subjects had three minutes of recovery after each emotional trigger. Participants self-reported their stress status before, during the delivery of the corresponding emotion trigger and after the last trigger, the questions were answered progressively. The closed questions that were formulated during the experiment were in a 7-point-ordinal-scale (presented on-screen). For instance, delivering first an environmental trigger and then a cognitive trigger, the sequence of questions were as follows (see Figure 5.2):

- **Q_1:** How stressed are you at this moment?
- **Q_2:** How stressed were you WHILE listening the noise?
- **Q_3:** How stressed are you at this moment?
- **Q_4:** How stressed were you WHILE doing the color task?
- **Q_5:** How stressed are you at this moment?

5.4 Data collection

The twelve subjects S01-S12 interacted with two emotional triggers successfully (i.e. environmental and cognitive). The experiment obtained an ethical approval from the

Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science of the University of Twente. Raw data and questionnaires answers were encrypted (WinZip AES encryption: 128-bit AES) and stored in a secured remote location for later analysis.

The experimental design was validated with a pilot study involving two participants (who did not take part in the final evaluation), to ensure the task descriptions were fully understandable, its implementation error-free, and to check the time and any further issue regarding the experimental design. The initial experiment was originally designed with three emotional triggers (cognitive, social and environmental), and approximately it lasted forty minutes. The feedback collected in the pilot suggested that the experiment was going to last too long and that the social trigger was not causing stress as expected given the time available. Hence, this design was changed to exclude the social-emotional trigger in order to prevent issues and reduce the experiment time approximately to thirty minutes (around 6480 EDA samples).

Table 5.1: Labeled results of the questionnaires and stress detector.

Subject	**Trigger**	**Reported stress**	**Stress detector**	**Trigger**	**Reported stress**	**Stress detector**
S01	Cognitive	Stressed	Stressed	Environmental	Not stressed	Not stressed
S02		Not stressed	Not stressed		Not stressed	Not stressed
S03		Not stressed	Not stressed		Stressed	Stressed
S04		Not stressed	Not stressed		Not stressed	Not stressed
S05		Stressed	Not stressed		Stressed	Not stressed
S06		Stressed	Not stressed		Not stressed	Not stressed
S07		Not stressed	Not stressed		Not stressed	Not stressed
S08		Not stressed	Not stressed		Not stressed	Not stressed
S09		Not stressed	Not stressed		Not stressed	Not stressed
S10		Stressed	Stressed		Not stressed	Not stressed
S11		Not stressed	Not stressed		Not stressed	Stressed
S12		Not stressed	Not stressed		Not stressed	Stressed

5.5 Threats to validity

5.5.1 Internal validity.

As the main objective is to evaluate the performance of the implemented stress detector, three well-known emotional triggers from the psychology community were selected for this study. However, given that these triggers had not been used previously in Software Engineering, the selected triggers could not always generate stress on the subjects (programmers and researchers in computer science) due to different other factors (*e.g.,* greater resilience) that were not investigated in this study. Given this interaction with different emotional triggers (treatments), with the purpose of avoiding that the the first emotional trigger does not affect on the next one, periods of relaxing and recovery were set out.

Another possible threat is the effect of the instrumentation used during the experiment (*i.e.,* E4-Wristband), which could also have been causing any stress level. To know whether this instrument could be considered as additional potential triggers, participants filled a post-questionnaire regarding this issue for further investigation.

5.5.2 External validity.

Given the low number of subjects and the fact they were researchers working in computing-related areas but not fully working as software engineers, one potential threat to external validity is regarding the generalization of these results. Moreover, as the controlled experiment was conducted in a lab setting, involving real practitioners would have been harder. This lab-setting still allows evaluating the stress detector without interruption of external factors (*i.e.,* meetings, calls) as a first necessary step to validate and continue developing the detector.

5.5.3 Construct validity.

The use of a single device to measure physiological stress (a construct) could be considered as main threat to construct validity of this study.

5.6 Results

The self-reported stress was rated in 7-point ordinal-scale questions that were gathered before and during the experiment (1 = "*not stressed*" to 7 ="*extremely*"). The overall self-reported score was labeled as "*stressed*" when the difference led to an increase equal o higher than 3 (threshold) in the perceived value of self-reported stress; otherwise, it

was labeled as "*not stressed*" according to (Bogomolov et al., 2014). Table 5.1 summarizes the labels for each subject to assess the accuracy or *trueness* of the stress detector, by comparing the computed label with the final self-reported label; red cells indicate the cases where the stress detector missed[2].

For answering our research question, the following (well-known) metrics were selected regarding: precision (Equation 5.1), recall (Equation 5.2) and accuracy (Equation 5.3):

$$precision = \frac{TP}{TP + FP} \quad (5.1) \qquad recall = \frac{TP}{TP + FN} \quad (5.2)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.3)$$

where TP indicates true positives, TN true negatives, FP false positive and FN false negatives. In this case, examples where reported stress and stress detector are labeled as *stressed* are considered as true positive (see Table 5.1).

The obtained results are: an accuracy of 79.17%, a precision of 60% and a recall of 50%. These results show a good accuracy in comparison with other machine-learning based on recognition methods, because it oscillates between 70% and 85% (Garcia-Ceja et al., 2016; Alberdi et al., 2016), values reported in the literature of stress recognition using physiological data. It is also important to remark that most of these existing recognition methods do not report precision and recall measures. Overall, a method that gets a high recall value it is considered as a good detector. However, the recall of the stress detector is low (50%) due to the threats that were identified after the results analysis.

5.7 Chapter discussion

This chapter presents the stress detection process of the arousal-based statistical method. The different algorithms and techniques used for supporting such detection process were implemented and integrated as part of the real-time stress detector, in a single processing pipeline. In order to evaluate its accuracy, an experiment was conducted with 12 subjects using the E4-wristband device to gather physiological data. Comparing the outcome of the stress detector with the reported by each subject (perceive stress), the detector obtained an accuracy of 79.17%.

An interesting observation is that although some subjects did not feel stress when an emotional trigger was delivered, the outcome of the detector was consistent with the

[2]Raw data and details of each subject can be found at https://goo.gl/e5Xtg2

corresponding perceived stress value. However, from this observation, the emotional trigger was not very effective for generating stress in all cases (subjects). A possible explanation for this might be due to the different resilience extent of the participants or the need to exposing them longer to the stimuli. As a future empirical work, beyond checking the accuracy of the detector, more research is needed regarding the role of emotional triggers and resilience of people working in office workplaces (*e.g.,* developer facing stress in unexpected situations).

Chapter 6

Evaluation of the actionable emotion detector

This chapter presents the design and results of the experiment for evaluating the performance of the actionable emotion detector. This experiment was carried out in the own working environment of each subject, and all participants were exposed to three emotional triggers and one software service. The first three sections detail the design of the experiment, including subjects, instrumentation, and procedure. Finally, the results and discussion of this chapter are presented in the last two sections.

6.1 Design of the experiment

The goal of this experiment is to evaluate the performance of the actionable emotion detector in semi-controlled conditions. The test was performed in the own environment of each subject (*i.e.,* office workers) using different stimulus that could cause stress (*i.e.,* emotional triggers and software service). From this goal and the proposed requirements, the following research questions are inferred:

- **RQ_1:** How accurately is the context analyzer able to recognize the different emotional triggers?
- **RQ_2:** Do the context analyzer, emotion detector and inference engine enable to recognize actionable emotions when the user interact with software services and different types of emotional triggers?
- **RQ_3:** What is the adequate temporal interval needed for detecting actionable emotion states?

The use of emotional triggers in the experiment is crucial for stimulating stress on the subjects. The delivery of persuasive messages (service) could also act as an

emotional trigger if the service caused a negative emotion on the user due to its not adequate configuration (e.g. a higher volume). Therefore, with the purpose of verifying whether the messages delivered by the mobile application cause any user dissatisfaction, two additional research questions were formulated as follows:

- **RQ_4:** How is the user satisfaction affected by the delivery of the software service (persuasive messages)?
- **RQ_5:** How comfortable is the interaction with the E4-wristband?

The experiment involved eight volunteer subjects from the University of Twente. For instance, Figure 6.1 shows a subject of this experiment working in her office. They were researchers in computing areas (*e.g.,* master students, Ph.D. candidates), their ages ranged between 23 and 31 years old (five men and three women).

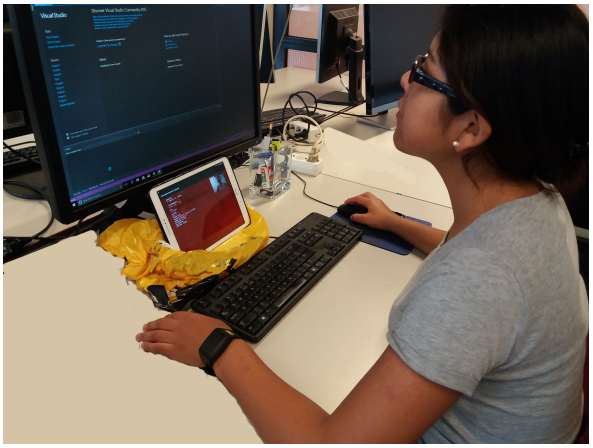


Figure 6.1: A subject in her own working environment and the actionable emotion detector at run-time.

6.2 Instrumentation and procedure

The experiment was carried out in the subject working environment; basically, all subjects' rooms were equipped with a table, a computer, and a chair. The actionable emotion detector was installed on a tablet that was placed in front of the participant to do not interrupt his activities, and, at the same time to collect data using the device camera as shown in Figure 6.1.

The evaluation followed the *within-subject design*, *i.e.,* the subjects were exposed to three emotional triggers and a software service. The Health Care Reminder App (a pill reminder), which sends persuasive messages in different persuasiveness level (Suni Lopez and Condori-Fernandez, 2017), was also installed on the tablet; it was used as the software service in this experiment. Additionally, the configuration of the

Health Care Reminder was intentionally changed to generate stress on the subject (*i.e.,* number of repetitions, level of volume and the persuasiveness level); Table 6.1 shows the configuration for each delivered message during this experiment.

Table 6.1: Configuration of the delivered messages during the experiment.

	Persuasive message	**Volume**	**#Repetitions**	**Level**
Message_1	"*It is time to take your pill. Your health could worsen if you do not take your medicines.*"	50%	1	1 (Scarcity)
Message_2	"*It is time to take your pill. You should take your pill remember that you committed to improve your health.*"	75%	3	3 (Commitment)
Message_3	"*It is time to take your pill. You must take your pill for your well-being.*"	100%	5	4 (Authority)

All participants received three persuasive messages in the same order, and they were exposed to three different emotional triggers (each one of five minutes) in random order: i) environmental noise, where participants listen to fire alarm sounds by five minutes; ii) physical movements, subjects are asked to walk by five minutes around the working place, and iii) interacting with others, participants were interrupted by the experiment leader with a casual conversation. The experiment lasted about 100 minutes and consisted of two phases as shown in Figure 6.2.

In phase 1, the subjects were asked to read and sign the consent form, which describes the purpose and the procedure of the experiment. Subjects were informed beforehand about the sensing device and the possibility of experiencing some episodes of stress during the experiment. Furthermore, they were informed that during this experiment they could continue with their work with the limitation that they cannot go out of their room. After signing the consent form, each subject got put on and adjusted the E4-wristband to enable the gathering of physiological data; they also were asked to complete a demographic questionnaire. This phase lasted around five minutes.

In phase 2, the first 15 minutes were used to get a baseline, *i.e.,* subjects were asked to stay quiet and relaxed. Then, as it is depicted in Figure 6.2, the subjects started to interact with the emotional triggers in the minutes 15, 45 and 75, for five minutes. Also, the persuasive messages (from the software service), which were detailed in Table 6.1, were delivered in three moments:

- $Message_1$ at minute 30, the Health-Care Reminder app sent the first message at level 1 with a volume of 50% without repetitions.
- $Message_2$ at minute 60, the Health-Care Reminder app sent the second message at level 3 with 75% of volume, and the same message was repeated three times

with 5 seconds between each message.

- *Message*$_3$ at minute 90, the third message at level 4 with maximum volume is delivered, and the same message was repeated five times with 5 seconds between each message.

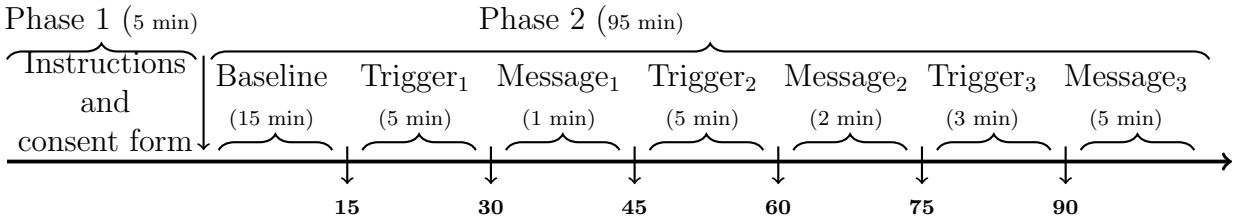


Figure 6.2: Experiment procedure and timeline: this experiment last 100 minutes and has two phases. In phase 1, the subjects were instructed about the experiment. In phase 2, the participants were exposed by five minutes to three emotional triggers (*i.e.,* environmental noise, physical movements and interaction with others) in different intervals of time. They also received three persuasive messages from the Health-Care Reminder App at three different times.

Finally, each participant answered a post-experiment questionnaire in Likert-scale about his stress status and comfortability in different parts of our experiment. The questions of the survey and the answers are shown in Table 6.2.

6.3 Results

The context analyzer is a central component in the actionable emotion detection because it provides relevant information to the inference engine of the possible threats of the user context for the evaluation of the actionable emotions; in this direction, the RQ$_1$ evaluates the performance of the algorithms for interpreting data from the context analyzer.

With the information provided by the context analyzer and the emotion detector, the RQ$_2$ examines the performance of the inference engine for recognizing actionable emotions. In this context, Table 6.2 summarizes the answers of the subjects during the post-questionnaire[1], where the first three questions collect the self-report of the perceived stress of each message for the assessment of the accuracy of the actionable emotion detector. Additionally, the RQ$_3$ analyzes the adequate temporal interval for detecting actionable emotions regarding the freshness and accuracy requirements.

The last three questions collect information about the comfortability and user satisfaction with the treatments used in the experiment (RQ$_4$ and RQ$_5$).

[1]Raw data and details for each subject can be found at https://goo.gl/y8VSYc

Table 6.2: Answers of the subjects for the post-experiment questionnaire.

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree	Total Answers
1. I felt stress with the first persuasive message.	-	25%	25%	25%	25%	8
2. I felt stress with the second persuasive message.	-	62.5%	-	37.5%	-	8
3. I felt stress with the third persuasive message.	12.5%	62.5%	-	12.5%	12.5%	8
4. I felt comfortable using the E4-Wristband.	12.5%	-	37.5%	50%	-	8
5. The repeated messages caused me stress.	12.5%	12.5%	25%	50%	-	8
6. I understood very well each message delivered by the App.	-	-	75%	25%	-	8

6.3.1 RQ_1: How accurately is the context analyzer able to recognize the different types of emotional triggers?

In this experiment, the context analyzer was integrated with three methods for processing the raw data gathered from three sensors (*i.e.,* the accelerometer of the E4-wristband, the microphone and the camera of the mobile phone) and detecting the corresponding emotional triggers (*i.e.,* physical movements, environmental noise and interaction with others).

Table 6.3 shows the results of the context analyzer for detecting each emotional trigger, and the *Stress Detector* column refers to stress state recognition in the interval of time when an emotional trigger was delivered. The context analyzer detected all episodes about the environmental noise trigger (with an accuracy of 100%). For the physical movement triggers, the context analyzer also got an accuracy of 100%. Nevertheless, since the accelerometer sensor was placed on the subject's wrist, when the subject moved his hand rapidly the sensor provided false positives of movements episodes. On the other hand, for detecting the trigger named interaction with others, the context analyzer used a face recognition algorithm with the frontal camera of the

mobile device. Since the algorithm needs that subjects show their face in the scene captured by the camera; in many situations, the method lost the tracking of subjects since participants left the scene. For those reasons, the context analyzer obtained a low accuracy of 37.5% for detecting this trigger.

Table 6.3: Results of the emotional trigger detection. The "context analyzer" column indicates if the algorithm used by the context analyzer detected the emotional trigger. The "stress detector" column refers if the stress detector recognized a stress state when the emotional trigger was delivered.

	Environmental noise		**Physical movements**		**Interaction with others**	
Subject	**Context analyzer**	**Stress Detector**	**Context analyzer**	**Stress Detector**	**Context analyzer**	**Stress Detector**
S01	Detected	Stressed	Detected	Not-Stressed	Not-Detected	Not-Stressed
S02	Detected	Not-Stressed	Detected	Not-Stressed	Not-detected	Not-Stressed
S03	Detected	Not-Stressed	Detected	Not-Stressed	Yes	Not-Stressed
S04	Detected	Not-Stressed	Detected	Not-Stressed	Not-Detected	Not-Stressed
S05	Detected	Not-Stressed	Detected	Stressed	Yes	Not-Stressed
S06	Detected	Not-Stressed	Detected	Not-Stressed	Not-Detected	Not-Stressed
S07	Detected	Not-Stressed	Detected	Not-Stressed	Yes	Not-Stressed
S08	Detected	Not-Stressed	Detected	Not-Stressed	Not-Detected	Not-Stressed

6.3.2 RQ_2: Do the context analyzer, emotion detector and inference engine enable to recognize actionable emotions when the user interact with software services and different types of emotional triggers?

In order to investigate the performance of the actionable emotion detector in terms of its accuracy, recall and precision, firstly the collected data through the self-report questionnaire was normalized to a binary option. The self-reported score is labeled as *Stressed* when the answers are *Agree* or *Strongly Agree*; otherwise, it is labeled as *Not stressed*. Table 6.4 summarizes the normalized answers of all participants, by comparing with the result of the detector to assess its accuracy; cells colored as red indicate the cases where the stress detector output does not match exactly with the self-reported state of stress.

Overall, for the assessment of the actionable emotion, the context analyzer did not detect any emotional trigger in the moments where the service was delivered; this

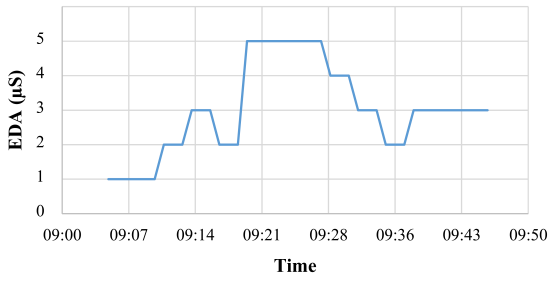
information was a fundamental input to the inference engine, which for determining what detected emotion is actionable, it mainly used the information of the emotion detector (stress detector). Therefore, the analysis of the performance of the actionable emotion was based on comparing the self-reported state of stress and the output of the inference engine. After the analysis, the actionable emotion detector obtained an accuracy of 75%, recall of 45% and precision of 80%.

Table 6.4: Labeled results of the questionnaires and stress detector.

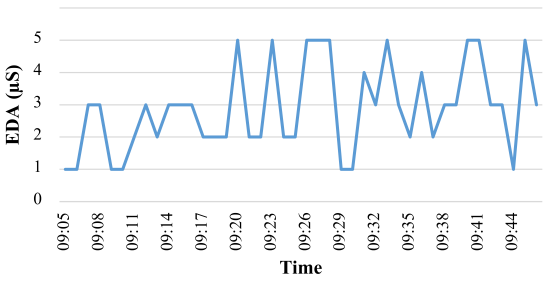
	Message$_1$		**Message$_2$**		**Message$_3$**	
Subject	**Reported stress**	**Stress detector**	**Reported stress**	**Stress detector**	**Reported stress**	**Stress detector**
S01	Not stressed	Not stressed	Not stressed	Not stressed	Not stressed	Not stressed
S02	Not stressed	Not stressed	Not stressed	Not stressed	Not stressed	Not stressed
S03	Stressed	Stressed	Not stressed	Not stressed	Not stressed	Not stressed
S04	Stressed	Not stressed	Stressed	Stressed	Not stressed	Not stressed
S05	Not stressed	Stressed	Stressed	Not stressed	Stressed	Not stressed
S06	Stressed	Not stressed	Not stressed	Not stressed	Not stressed	Not stressed
S07	Stressed	Stressed	Stressed	Stressed	Not stressed	Not stressed
S08	Not stressed	Not stressed	Not stressed	Not stressed	Stressed	Not stressed

6.3.3 RQ$_3$: What is the adequate temporal interval needed for detecting actionable emotion states?

This interval of time is mainly determined by the algorithm of the emotion detector (the arousal-based statistical algorithm). Then, when the interval of time is reduced (*e.g.,* one minute), the assessment of the stress got many false positives of stress states, as is shown in Figure 6.3b, and consequently it obtained a low accuracy of 50%. In contrast, when the interval of time is increased (to five minutes), the algorithm loses more episodes of stress (with 37.5% of accuracy). In general, an adequate temporal interval was determined in three minutes, as is presented in Figure 6.3a, which also obtained a balanced accuracy of 75% in this experiment and 79.17% in the previous experiment described in Chapter 5.



(a) Assessment of the stress using three minutes as interval of time



(b) Assessment of the stress using one minutes as interval of time.

Figure 6.3: Assessment of the stress using different intervals of time.

6.3.4 RQ$_4$: How is the user satisfaction affected by the delivery of the software service (persuasive messages)?

Two questions in the post-questionnaire (Q5 and Q6 in Table 6.2) were used to analyze user satisfaction. Figure 6.4 presents the number of answers for the question Q5: *The repeated messages caused me stress.* A moderate trend to *Agree* means the delivered messages caused certain discomfort. This result may be explained by the fact that the configuration for the messages was intentionally changed to generate this feeling on subjects.

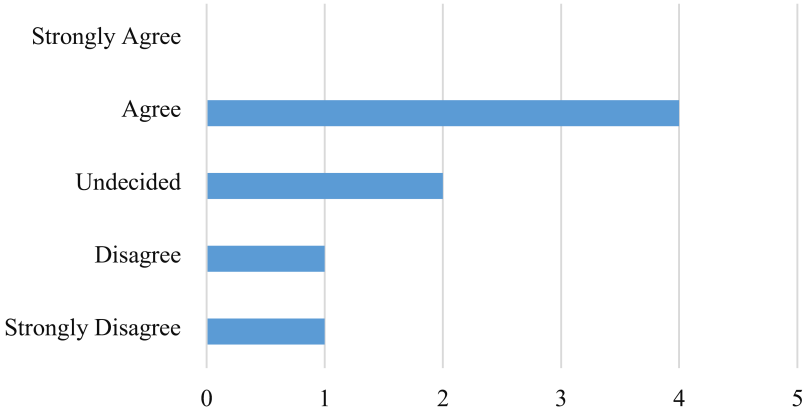


Figure 6.4: Trend of answers of question: *The repeated messages caused me stress.*

Q6: *I understood very well each message delivered by the App.* As shown in Figure 6.5, exists a clear trend to *Agree and strongly agree* options. This means the configuration of the voice assistant was good enough (*i.e.,* velocity of pronunciation) to understand the delivered messages.

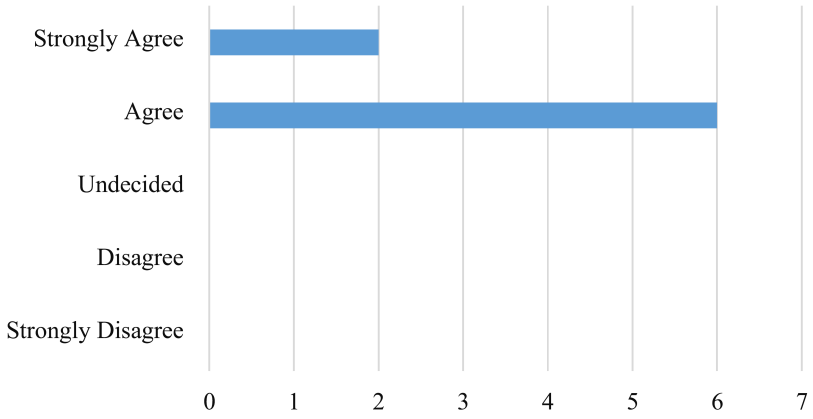


Figure 6.5: Trend of answers of question: *I understood very well each message delivered by the App.*

Additionally, the *Stress Detector* column in Table 6.3 refers that in two user the environmental noise and the physical movements cause them stress states.

6.3.5 RQ_5: How comfortable is the interaction with the E4-wristband?

Figure 6.6 shows that 50% of the subjects were comfortable using the wearable device. Surprisingly, a considerable amount of respondents were undecided about the comfortability of using the E4-wristband due to privacy concerns. This privacy issue requires more research that is not part of this study. On the other hand, it is also important to remark that only one subject disagreed with the comfortability feature of the E4-Wristband.

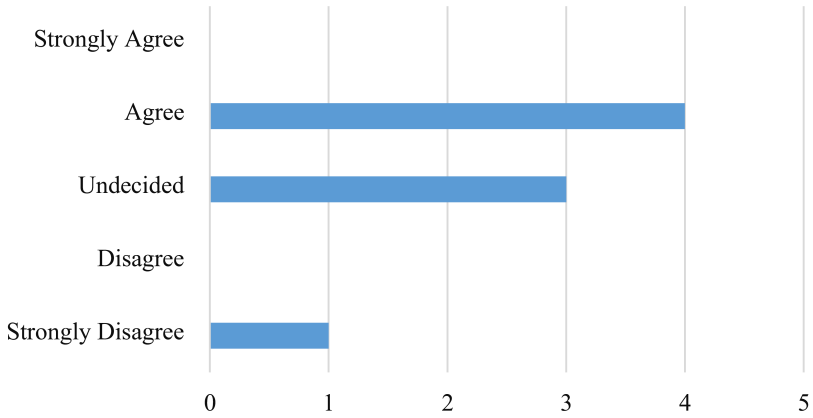


Figure 6.6: Trend of answers of question: *I felt comfortable using the E4-Wristband.*

6.4 Chapter discussion

This chapter presented an experiment to evaluate the performance of the actionable emotion detector in semi-controlled conditions. In general, the actionable emotion detector obtained an accuracy of 75%; the context analyzer was integrated by three methods: the environmental noise and physical movements evaluators obtained an accuracy of 100%, but the analyzer of interaction with others got an accuracy of 37.5% (RQ_1 and RQ_2). As there not exist related works for comparing our work, the obtained results are established for future works in this field. The RQ_3 contributed to address one challenges proposed in the actionable emotion detection componet of the HAPPYNESS framework, regarding the freshness requirement. This challenge is related to determine the adequate temporal interval needed for detecting actionable emotion states (Condori-Fernandez, 2017). In this context, thanks the data collected from the experiment, the interval of time required for detecting actionable emotions was determined in three minutes, this time allowed to get good results in accuracy but also in freshness. The research has also shown that user satisfaction might be affected by changes in the configuration of delivered reminder messages (RQ_4). These configurations were intentionally manipulated respect to the volume, frequency and persuasiveness level. As it was not possible to verify the comfortability of E4-wristband (RQ_5); it is unknown if this kind of wearable devices could become accepted by users. A further study could investigate other variables that could have been affecting to the non comfortability felt by the subjects (*e.g.,* privacy).

Overall, using the selected emotional triggers and available sensors, the actionable emotion detector has a good performance concerning freshness and accuracy. It is important to remark that both requirements are met when the interval of time is of three minutes, which was determined by the algorithm of the emotion detector (the arousal-based statistical approach). As a future work, more research is needed regarding reducing this interval with others algorithm for recognizing physiological stress. As the first experiment (reported in Chapter 5) and this second experiment were carried out in controlled and semi-controlled conditions, in the following Chapter, a case study is presented with the objective of corroborating whether the performance of the actionable emotion detector is still good in non-controlled conditions.

Chapter 7

Evaluation of actionable emotion detector in the wild

This Chapter presents the case study of a tourist traveler for analyzing the performance of the actionable emotion detector in the wild — *i.e.,* real-life environments where the user is exposed to a diverse set of stimuli under uncontrolled conditions such as diverse backgrounds (indoor/outdoor).

In particular, this case study focuses on a tourist, whose physiological data is monitored during a four-day visit to Rome (Italy), and it was used as inputs for our actionable emotion detector. The subject interacted with a persuasive mobile application (Health Care Reminder), which sends persuasive messages for reminding the intake time of the tourist's pills. EDA signals, skin temperature, and physical activity data were monitored using the E4-wristband. These data were then combined with post hoc interviews, including time, locations, and activities to aid interpretation of the obtained results.

This Chapter is organized as follows: Section 7.1 presents the used methods, description of the subject, treatment of the study, and measurement of data. The description of our findings is presented in Section 7.2. Finally, Section 7.3 discusses the findings of the case study.

7.1 Method

In order to explore the performance of the actionable emotion detector in real-life conditions, this study was conducted using mixed methods following the single subject design - post hoc interviews and a four-day diary study of a tourist.

Post hoc interviews. An oral interview was conducted immediately after each message (from the reminder app) was delivered to the subject (tourist), asking about

his stress perception. The survey included only one yes/no question: *Did you feel stress while the reminder was playing?*.

Dairy study. During the four days, details about the activities, places, and events that were realized by the subject were stored. Additionally, the Rome weather was saved [1] to aid the interpretation of stress states. From this goal and the proposed requirements, the following research questions are inferred:

- **RQ$_1$:** How do emotional triggers from the user context affect the detection of the actionable emotions in real-life conditions?
- **RQ$_2$:** In which kind of tourist activities the user tend to feel more stress?
- **RQ$_3$:** How accurately is the actionable emotion detector able to recognize actionable emotions in real-life conditions?

7.1.1 Analysis unit

The persona method was applied for identifying and capturing significant details of the subject. The details considered in a persona template include personal information, goals, preferences, challenges, and skills. User characteristics are based on data gathered in the "real world". In this context, Adlin and Pruitt (2010) provides explicit steps to guide the process of creating a persona.

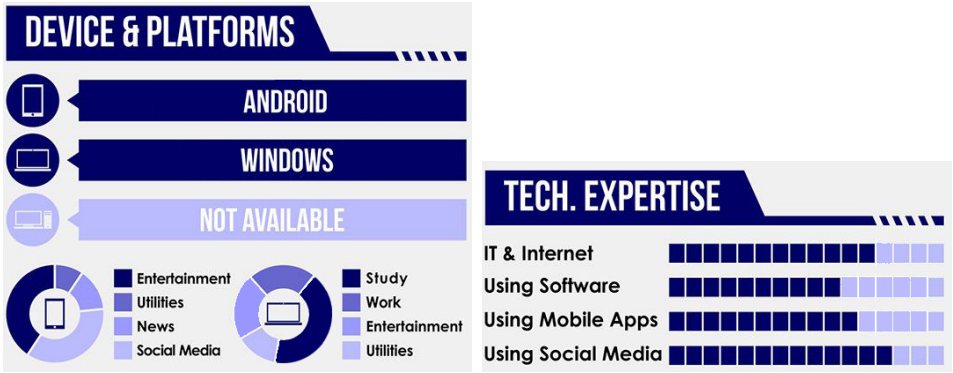


Figure 7.1: Technological skills of the subject according to persona method.

The subject is a Peruvian math teacher. He is 49 years old, his native language is Spanish, and he does not speak another language. He was diagnosed with *non-allergic rhinitis* that affects the nasal mucus and produces sneezing, itching, obstruction, nasal secretions, and sometimes lack of smell (Nozad et al., 2010). For that reason, he was prescribed with oral antihistamines (cetirizine 10 mg) twice by day (one pill in the morning and other in the night).

[1] AccuWeather website: https://www.accuweather.com/en/it/rome/213490/month/213490?monyr=8/01/2017